2011

Homologous Pairing Through Dna Driven Harmonics-- A Simulation Investigation

Richard J. Calloway
University of Central Florida

HOMOLOGOUS PAIRING THROUGH DNA DRIVEN HARMONICS – A SIMULATION INVESTIGATION

by

RICHARD J. CALLOWAY
B.S.E. University of Central Florida, 1987
M.B.A. University of Central Florida, 2000

A dissertation to be submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Modeling and Simulation
in the College of Engineering and Computer Science
at the University of Central Florida
Orlando, Florida

Fall Term
2011

Major Professor: Michael Proctor

ABSTRACT

The objective of this research is to determine if a better understanding of the "molecule of life", deoxyribonucleic acid or DNA can be obtained through Molecular Dynamics (MD) modeling and simulation (M&S) using contemporary MD M&S. It is difficult to overstate the significance of the DNA molecule. The now-completed Human Genome Project stands out as the most significant testimony yet to the importance of understanding DNA. The Human Genome Project (HGP) enumerated many areas of application of genomic research including molecular medicine, energy sources, environmental applications, agriculture and livestock breeding to name just a few. (Science, 2008) In addition to the fact that DNA contains the informational blueprints for all life, it also exhibits other remarkable characteristics most of which are either poorly understood or remain complete mysteries.

One of those completely mysterious characteristics is the ability of DNA molecules to spontaneously segregate with other DNA molecules of similar sequence. This ability has been observed for years in living organisms and is known as "homologous pairing." It is completely reproducible in a laboratory and defies explanation. *What is the underlying mechanism that facilitates long-range attraction between 2 double-helix DNA molecules containing similar nucleotide sequences?* The fact that we cannot answer this question indicates we are missing a fundamental piece of information concerning the DNA bio-molecule. The research proposed herein investigated using the Nano-scale Molecular Dynamics NAMD (Phillips et al., 2005) simulator the following hypotheses:

$\mathbf{H}_{(\text{Simulate Observed Closure NULL})}$ **: =** Current MD force field models when used to model DNA molecule segments, contain sufficient variable terms and parameters to describe and reproduce

directed segregating movement (closure of the segments) as previously observed by the Imperial College team between two Phi X 174 DNA molecules. $\mathbf{H}_{(\text{Resonance NULL})}$ **: =** Current MD force field models when used to model DNA molecule segments in a condensed phased solvent contain sufficient variable terms and parameters to reproduce theorized molecular resonation in the form of frequency content found in water between the segments.

$\mathbf{H}_{(\text{Harmonized Resonance NULL})}$ **: =** Current MD force field models of DNA molecule segments in a condensed phase solvent produce theorized molecular resonation in the form of frequency content above and beyond the expected normal frequency levels found in water between the segments.

$\mathbf{H}_{(\text{Sequence Relationship NULL}):}$ = The specific frequencies and amplitudes of the harmonized resonance postulated in $\mathbf{H}_{(\text{Harmonized Resonance NULL})}$ are a direct function of DNA nucleotide sequence.

$\mathbf{H}_{(\text{Resonance Causes Closure NULL})}$ **: =** Interacting harmonized resonation produces an aggregate force between the 2 macro-molecule segments resulting in simulation of the same directed motion and segment closure as observed by the Imperial College team between two Phi X 174 DNA molecules.

After nearly six months of molecular dynamic simulation for H(Simulate Observed Closure NULL) and H(Resonance Causes Closure NULL) no evidence of closure between two similar sequenced DNA segments was found. There exist several contributing factors that potentially affected this result that are described in detail in the Results section. Simulations investigating H(Resonance NULL), H(Harmonized Resonance NULL) and the emergent hypothesis H(Sequence Relationship NULL) on the other hand, revealed a rich selection of periodic pressure variation occurring in the solvent between simulated DNA molecules. About

20% of the power in Fourier coefficients returned by Fast Fourier Transforms performed on the pressure data was characterized as statistically significant and was located in less than 2% of the coefficients by count. This unexpected result occurred consistently in 5 different system configurations with considerable system-to-system variation in both frequency and magnitude. After careful analysis given the extent of our experiments the data was found to be in support of H(Resonance NULL), and H(Harmonized Resonance NULL) . Regarding the emergent hypothesis H(Sequence Relationship NULL), further analysis was done on the aggregate data set looking for correlation between nucleotide sequence and frequency/magnitude. Some of the results may be related to sequence but were insufficient to prove it. Overall the conflicting results were inconclusive so the hypothesis was neither accepted nor rejected. Of particular interest to future researchers it was noted that the computational simulations performed herein were NOT able to reproduce what we know actually happens in a laboratory environment. DNA segregation known to occur in-vitro during the Imperial College investigation did not occur in our simulation. Until this discrepancy is resolved MM simulation should not as yet be considered a suitable tool for further investigation of Homologous Chromosome Pairing. In Chapter 5 specific follow on research is described in priority of need addressing several new questions.

I dedicate this work to Jesus my Lord and Savior who carried me when I could go no further.

I dedicate this work to Donna my loving wife and friend since high school. Her emotional and financial sacrifices far outweigh my own because they were selfless and given generously while she raised our sons almost by herself. Without her by my side I would be nothing. I pray her patience with me will last for another 25 years of marriage.

I also dedicate this work to my wonderful sons Jason and Cameron who are growing into fine young men even though their Dad was not always there for them during the last 5 years of their lives.

Finally I dedicate this work to my Mom and Dad, Dick and Velma Calloway. For as long as I can remember, their love and confidence in me always gave me a positive attitude towards taking on tough challenges. It was easier to try because they were there for me. Their enthusiastic support helped give me the courage to take the first steps toward a doctorate. Although I lost both of them before I could complete the journey, I know somehow, together, they are rejoicing with me in my success. They knew I could do it.

ACKNOWLEDGMENTS

I owe my deepest gratitude to my advisor, committee chair and friend Dr. Michael Proctor. His straight talk and sometimes painful to hear advice, successfully guided me through this complicated process. I am most grateful for his positive encouraging attitude that always made the challenges seem a little less formidable.

I also would like to acknowledge my mentor and friend Dr. John Sanford. His unique perspective of the world and heartfelt encouragement gave me the confidence I needed to boldly explore beyond all of my formal training.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

Chapter 1 Abstract:

Today there is a massive worldwide research effort on the DNA molecule moving forward at a furious pace. The results of this research are changing our civilization. The ability to determine the exact nucleotide sequence of a DNA molecule has not only impacted the world wide medical community but also the world's legal systems in ways we won't fully comprehend for generations.

DNA plays a truly unique role in the foundation of all life sciences as the repository for the blueprints of life. Its remarkable list of behaviors like precision replication, super coiling and automated self repair continue to draw steadily increasing attention from researchers in the physical sciences like chemistry and physics. DNA is single handedly responsible for the birth of molecular biology and its data storage capability and efficiency has inspired the growing new discipline of bioinformatics. DNA is at the center of scientific crossroads rich in fascinating high value problems for interested researchers.

Chapter 1 briefly discusses the history and significance of DNA research and presents a short review of its molecular structure. It is pointed out that the majority of recent research has a genomic focus even though enormous gaps still exist in our understanding of the molecule at the molecular level. One of the most intriguing behaviors of DNA, homologous pairing, is introduced identifying a key gap in our knowledge of the molecule. It is demonstrated that homologous pairing is DNA driven and a better understanding of the underlying mechanism is

likely to have a profound effect on the scientific community. The chapter closes with a discussion of plausible explanations and several implications.

History of General DNA research

Any discussion involving DNA or the history of genetics in particular should necessarily begin with Gregor Mendel. Between 1857 and 1865 Mendel carried out experiments with garden peas in which he established the foundational concepts of inheritance and heredity. He successfully published his research findings in 1866 to a summarily disinterested scientific community. Mendel was bitterly disappointed in the lack of interest in his research and did not continue his efforts beyond 1870. He died unrecognized for his accomplishments in 1884. It wasn't until sixteen years after his death that his work was re-discovered by biologists of the time and was re-published again in 1901. The work is now known as 'Mendel's Laws' and forms the basis of modern genetics. Although he had no concept of the DNA molecule itself Mendel is frequently referred to as "The Father of Genetics". He set the stage for subsequent research that would eventually identify the molecule actually responsible for heredity, DNA.

The earliest known research on the DNA molecule itself dates back to 1868 when the Swiss biologist Friedrich Miesher detected a substance from the nuclei of cells he called nuclein. We now know this substance contained DNA and histones. He successfully separated out nucleic acid from his nuclein and performed the first known study of DNA specifically. In 1889 a student of his, Richard Altman, named the substance "nucleic acid". Research progressed slowly for the next 40 years with no significant breakthroughs, although researchers began to suspect that nucleic acid was somehow linked to inheritance. Then in 1929 Phoebus Levene

identified the individual chemical components that comprise a DNA molecule. He correctly identified them as sugar, phosphate and 4 acid bases, adenine, cytosine, guanine and thymine. He believed that they were assembled in phosphate-sugar-base subunits that he called "nucleotides". He also correctly postulated that nucleotide units would string together connected by the phosphates making a backbone for a potentially long molecule. His theory was rejected by the scientific community until Watson and Crick conclusively determined the double-helix structure in 1953.

It was at about the same time in the early 1930's that molecular biology as a discipline was born. The birth of molecular biology was the result of research activity in the fields of biology, physics, chemistry and genetics converging on the structure and function of DNA. Hermann J. Muller recognized that X-rays caused genetic mutations in fruit flies (Drosophila) and began using this phenomenon to investigate the size and structure of the gene. He eventually realized that as a geneticist he was limited in just how much he could learn from mutagenic results obtained by bombarding fruit flies with X-rays. (Muller, 1927) In his 1936 essay "Physics in the Attack on the Fundamental Problems of Genetics" Muller stated:

> "The geneticist himself is helpless to analyze these properties further. Here the physicist as well as the chemist must step in. Who will volunteer to do so?" (Muller, 1937)

During the next 20 years the research community responded with keen interest in DNA molecular structure from esteemed physicists like Erwin Schrödinger and Max Delbrük. Schrödinger applied the principles of quantum physics to attempt to explain the stability as well as the mutagenic capabilities of genes. (Schrödinger, 1944) Delbrük was keenly interested in

how the separate disciplines of biology and physics might complement each other. In order to promote this mutual cooperation he subsequently formed "The Phage Group" in 1940 to highlight and facilitate collaboration between the fields. At about the same time Linus Pauling at the California Institute of Technology was studying large macromolecules from the perspective of structural chemistry. An exciting new analysis technique based on shining X-rays through materials of interest and measuring the diffraction angles on photographic plates, now known as X-Ray Crystallography, had reached atomic resolution accuracy by 1914 when Bragg analyzed the crystal structure of table salt. (Bragg, 1914) Pauling used the x-ray crystallography technique to build scale models of macromolecules and subsequently discovered the alpha-helical structure of protein. (Pauling, Corey, & Branson, 1951) Then in 1938 Warren Weaver, Director of the National Sciences section of the Rockefeller Foundation, wrote a report to the foundation in which he stated:

> "And gradually there is coming into being a new branch of science – molecular biology – which is beginning to uncover many secrets concerning the ultimate units of the living cell….in which delicate modern techniques are being used to investigate ever more minute details of certain life processes." (quoted in Olby 1994, 442 The path to the double helix: the discovery of DNA revised edition.)

Weaver thus coined the term 'Molecular Biology'. Perhaps the most insightful explanation of the origin of the term was given by Francis Crick in 1965 when he stated:

> "I myself was forced to call myself a molecular biologist because when inquiring clergymen asked me what I did, I got tired of explaining that I was a mixture of crystallographer, biophysicist, biochemist, and geneticist." (Crick, 1965)

After 1938 there was only minor progress in the field until Rosalind Franklin and Maurice Wilkins succeeded in making DNA crystallize. This allowed them to use x-ray diffraction on DNA producing the first x-ray patterns. The now famous "photo 51" revealed for the first time the helix shape of the DNA molecule. Finally in 1953 James Watson and Francis Crick, admittedly inspired by "photo 51", accurately modeled the physical structure of DNA for which they and Maurice Wilkins received the Nobel Prize in 1962. Rosalind Franklin was not a recipient because she had died of ovarian cancer by that time.

The DNA Molecular Structure

Today we understand that the DNA molecular structure is rather simple in its most basic form. It is a polynucleotide, meaning it is a polymer whose monomer components are nucleotides. The basic nucleotide building blocks are 5-carbon sugars, a nitrogen base attached to the sugar, and a phosphate group. A DNA polymer is made up of 4 different nucleotides commonly denoted by their first letters A, G, C, and T corresponding to adenine, guanine, cytosine and thymine respectively. The structure of each base is illustrated below.

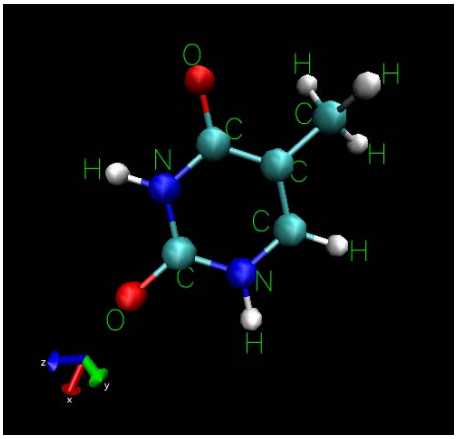


Figure 1: Thymine

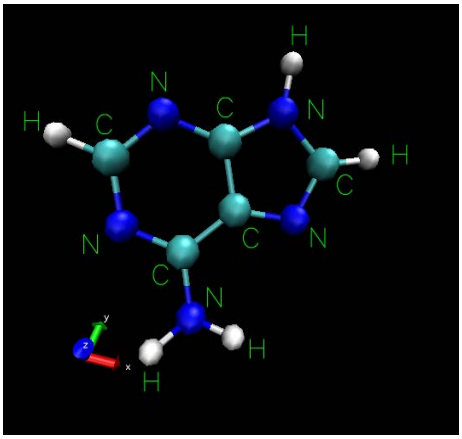


Figure 2: Adenine

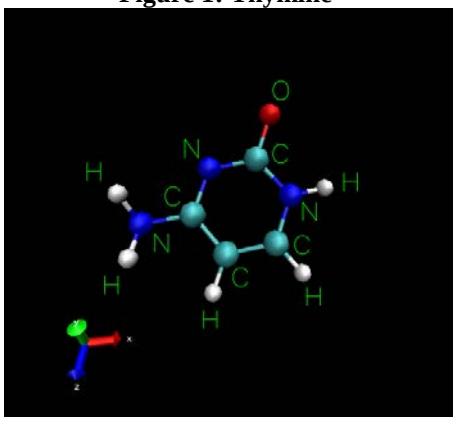


Figure 3: Cytosine

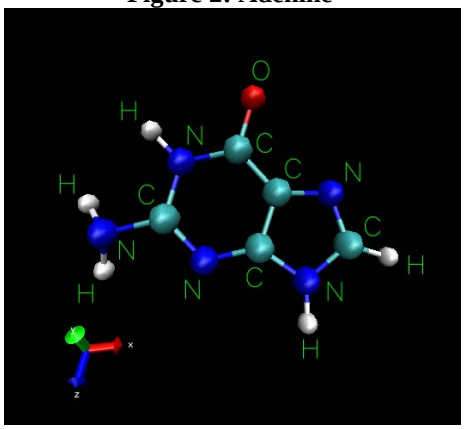


Figure 4: Guanine

The DNA backbone is a polymer chain made up of an alternating sequence of sugar and phosphate groups in a continuous sequence.

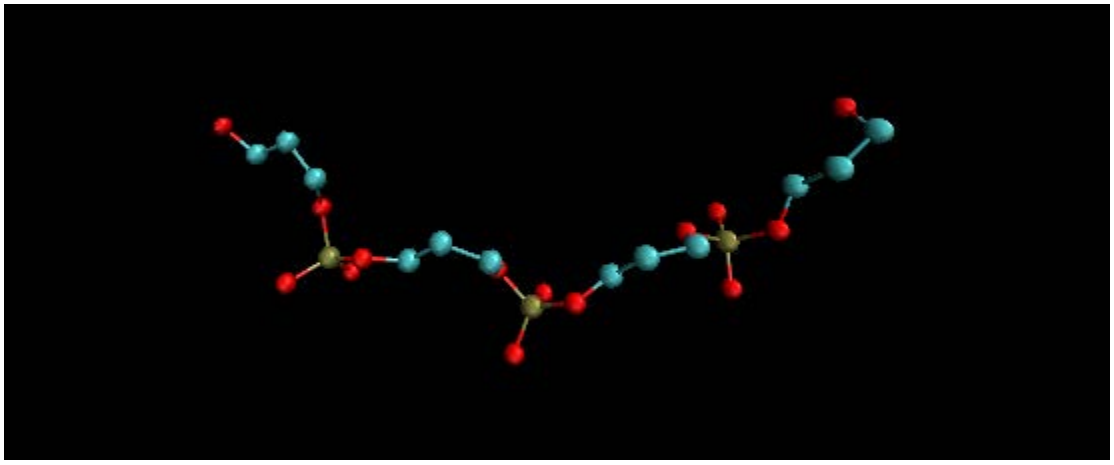


Figure 5: 4 Base Pair DNA Backbone Chain

The DNA molecule is composed of a long and varying sequence of A, G, C, and T bases attached to the chain and extending away from the backbone. The unbounded ends of the bases form the inside of a double helix when A forms 2 hydrogen bonds with T and G forms 3 hydrogen bonds with C. The 2 strands then form a helical right handed spiral where the planer bases stack on top of each other like steps in a spiral staircase. Below is the complete structure with 4 base pairs forming a classic double helix. The bases are color coded for illustrative purposes.

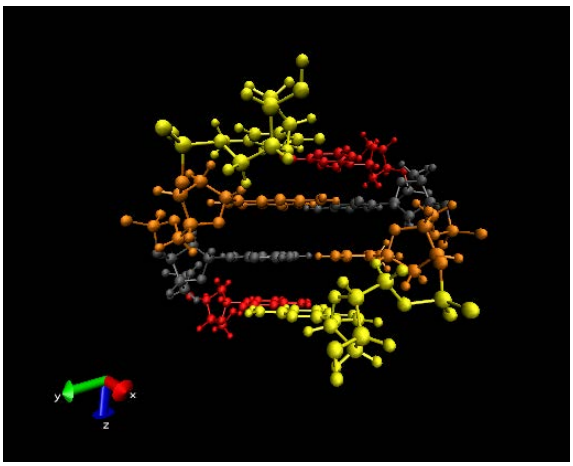


Figure 6: A G C T Polymer CPK Color View

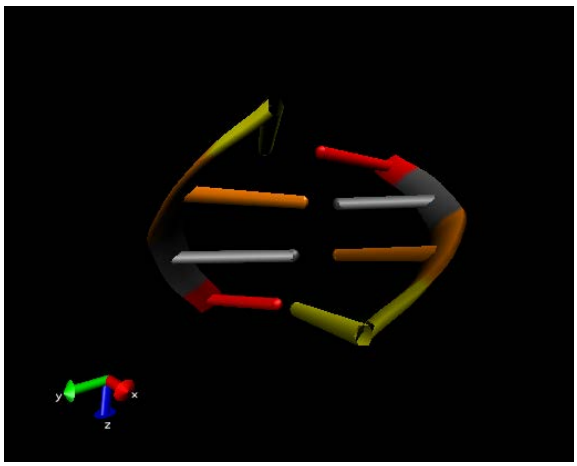


Figure 7: Ribbon View Showing Backbone

The molecule in Figure 6 contains exactly 252 atoms including hydrogens that bond to various open valences on the structure. Discussion of valence bond theory and molecular orbital theory, the two primary theories that govern chemical bonding are deferred until later. The helix has a diameter of about 2nm and is about 2nm long for a 4 base pair segment. This basic base-pair unit is then extended to a vast chain of nucleotides numbering in the millions forming a macromolecule. This very long molecule – much like a string of beads - is then wound around histone protein spheres about 11nm in diameter forming a structure much like thread around a spool (http://www.sci-ed-ga.org/modules/dna/anals/genedna.html).(Derhaag, 1996) . This string of beads is then packed into a denser string of histones increasing the diameter of the strand to 30 nm. This strand is further condensed and packed in a layered ribbon fashion forming a thicker rope like weave about 700nm in diameter. This rope weave then forms the familiar chromosome shape observed under a microscope.

It should now be apparent that from an atomic perspective a chromosome is massive and easily seen with an optical microscope. The smallest human chromosome (chromosome 21) is a single DNA segment that spans more than 47 million nucleotide base pairs. This is equivalent to about 3 X 10E9 atoms, far beyond today's computational modeling capability.

Genomic Focus

After the landmark discoveries of Crick and Watson were assimilated by the research community the nature of DNA research shifted from a structural molecular focus to a focus on information in the form of genes contained in the structure. Genes form the basic biological information code of organisms. The sum total of all the genetic information contained in an

organism is known as its genome. Research efforts concentrated on correlating specific sections of chromosomes in various organisms with specific genes in a process generally known as 'mapping'. Little significant progress was made until the Human Genome Project officially began in October of 1990. The Human Genome Project was an international effort that ran from 1990 until 2003 that was intended to locate and identify what was thought to be 2,000,000 genes but only actually identified roughly 25,000 genes that make up the human genome. An additional project goal was to accurately sequence the more than 3 billion base pairs that comprise the genome as well. The project ended successfully in 2003 having met its stated objectives. The project directly resulted in more than 30 publications and numerous research spin-offs most of which are on-going today. In spite of a nearly 3 billion dollar price tag the countless benefits of this project will continue to materialize for many years to come.

Surprisingly, with so much attention and so many resources being recently devoted to genetic research, one of the greatest mysteries of the chromosome has received very little fanfare.

Homologous Pairing

This brings us to the primary focus of this research effort, long-range interaction between 2 double-helix DNA molecules containing similar nucleotide sequences, commonly referred to as homologous pairing. The definition of homologous chromosomes varies somewhat but in the most basic sense they are equal in length (number of base pairs), the same general shape, and contain the same amount of genetic information although the content of genetic material are not genetically identical. This means the base pair sequence of the DNA molecule comprising each

chromosome is only similar, not exact, with sequence variation correlating with intra-species characteristic (allele) variation. When 2 such similar chromosomes exist together in the same environment they are referred to as a "homologous pair" or HP. The phenomenon of "homologous pairing" was probably first noticed by Barbara McClintock in 1933 when she observed: "there is a tendency for chromosomes to associate 2-by-2 in the prophase of meiosis" (quoted by (Denise Zickler, 2006)). She was referring to the tendency for homologous chromosomes to "pair up" at a very early stage of meiotic prophase. This pairing is an essential pre-requisite to the process of genetic recombination and therefore essential to transmission genetics in general. (Sybenga, 1999) This pairing is not limited to meiosis. It also occurs in a variety of other biological processes where homologues are observed segregating including somatic and mitotic cells of Dipteran insects as well as vegetative examples. Sister chromatids also pair quite closely promoting the necessary segregation for both meiosis and mitosis. (McKee, 2004) With respect to homologous pairing, the literature remains sparse from the 1930's into the late 1990's. The topic appeared in publication again in an article by (D. Zickler & Kleckner, 1998) which is a thorough review of the leptotene-zygotene transition of meiosis. Although the review covers a broad variety of processes involved with meiosis it specifically identifies the molecular movement we're interested in and reports the complete lack of available data at the time of the writing. One year later, J. Sybenga published an article that is probably the first publication targeted specifically at this question titled "What makes homologous chromosomes find each other in meiosis? A review and an hypothesis" (Sybenga, 1999). In this paper Sybenga defines meiotic chromosome pairing as the long distance attraction between homologous sites, followed by aligning of chromosomal segments. As did Zickler and

Kleckner, Sybenga pointed out again that nothing at all is known about what brings homologous chromosomes together forming the synaptonemal complex. The author summarily states that "No satisfactory hypothesis has been presented for the biochemical and cell-biological processes involved". He concludes early-on that DNA-DNA interactions are at most only indirectly responsible for initial pairing and are insufficient to bridge the sometimes large distances between homologous chromosomes within a cell nucleus. He further maintains that double-stranded DNA is inefficient at recognizing homology and that single-stranded DNA long enough for long-distance recognition is not available in the nucleus because of attack by endonucleases. As a result of these pre-suppositions an hypothesis is formulated suggesting the existence of pairing proteins that form protein chains between DNA segments that mechanically pull segments together over long distances with a zipping or sliding action.

Another significant contribution to the mystery was made in 2003 by researchers at the University of California, Berkley. Observations of chromosome movement during meiotic bouquet formation were quantified in three dimensions. The observations revealed a gradual tightening of telomeres eventually forming a bouquet after about 6 hours. A computational simulation was devised in order to test whether random diffusion was sufficient for bouquet formation or if directed motion was at work. The two significant variables in their models were diffusion rate and directional bias. They adjusted these 2 variables over a wide range of values until they successfully matched the observed data. The results showed that non-random directed motion was required to reproduce the empirical data, implying that an active process was influencing chromosome movement toward the bouquet. (Carlton, Cowan, & Cande, 2003) The underlying mechanism was not identified by the study, only its likely existence.

The exact phenomenon appeared in publication again only a year later. In his paper "Homologous pairing and chromosome dynamics in meiosis and mitosis" Bruce McKee poses 2 insightful questions, First "Is there a single under-lying mechanism for pairing of homologous loci?" and second "Are there common mechanisms for linking sister chromatids and chromosomes in various segregation pathways?". The author noted that the literature abounds with observations and descriptions of pairing processes together with a correspondingly large number of hypothesized explanations, but the actual mechanisms in operation during homolog pairing remain completely unknown. (McKee, 2004) This effort served only to reiterate the question.

In 2005 Denise Zickler updates the scientific community with yet another review of the process from early homologue recognition to synaptonemal complex formation. She concludes "There is almost no understanding of the mechanistic basis for recombination-independent homologue recognition and juxtaposition…..Finally, the models for chromosome recognition and clustering into the bouquet discussed here are still at a highly speculative stage, underlining our ignorance, will hopefully shape future thinking and provoke new investigations".

Today, almost 75 years after Barbara McClintock first observed the phenomenon, we still have almost no understanding of the mechanisms by which these homologous chromosomes recognize each other, translate through the cytoplasm, and then precisely align at atomic resolution just prior to the formation of a synaptonemal complex. (Denise Zickler, 2006)

Homologous Pairing is DNA Driven

One bit of information that we do know for certain about the phenomenon is that the sequence of the nucleotide base pairs of a DNA molecule is the primary, if not the only variable, driving the pairing process. Let's examine this concept more closely. We already know that the effect of base pair sequence on a DNA molecule's structure topology and conformational dynamics is a critical factor in our understanding of the biochemistry of DNA. (Beveridge et al., 2004) The blatantly obvious importance of base pair sequence of course is that the sequence represents coded information that makes up the basic unit of heredity, the gene. Small variations in the sequences of individual gene segments make up alleles. An allele is a representative of a set of one or more gene variations within a species. A good example of a human allele is eye color. Blue eyes and green eyes may represent 2 alleles of the genes that control eye color. Through a process of transcription and translation a genetic sequence is expressed into a gene product like protein or RNA. These processes of DNA to RNA transcription followed by RNA to protein translation represent what is called the "central dogma" of biology.

For years researchers observing homologous pairing have suspected there were many more pieces to the puzzle then the processes encompassed by the central dogma. Today we have solid evidence that the sequence does in fact play a much more complex role than previously imagined. A recent study done at Imperial College London illuminated an amazing sequence effect that appears to be unrelated to short range chemical bonds and electrostatic forces. Up to now, the majority of hypotheses regarding chromosome pairing or chromosome movement within a cell have involved protein chains, association with the nuclear envelope (Carlton et al.,

2003), electrostatic forces or simple random Brownian movement. In a highly publicized news release from Imperial College London, findings were published that show conclusively that double-stranded DNA molecules of identical sequence will spontaneously segregate revealing homologous recognition and linear translation through at least 1nm of water in a protein free environment! (Baldwin et al., 2008)

Is Homologous Pairing simply the result of Intermolecular Forces?

To further this discussion it is necessary to standardize some terminology and define relationships between several inter-disciplinary concepts. At this point it becomes more expedient to categorize the atomic and molecular interactions we are looking at in a manner consistent with molecular modeling. Molecular modelers have logically categorized forces into 2 basic types, bonded and non-bonded even though they are all fundamentally electromagnetic forces which can be electrostatic or electrodynamic in nature (thus the potential for inter-disciplinary confusion). Bonded forces are defined as forces resulting only from covalent bonds. Non-bonded forces naturally are defined as forces resulting from all 'non-covalent' bonds i.e. electrostatic forces and van der Waals forces. Bonded or covalent forces are much stronger than non-bonded forces but they typically operate only within proximity of the outer electron shells of individual atoms, therefore they operate from between .075nm and .200nm. Homologous DNA molecules are not actually bonded together so the only known forces interacting between them are the non-bonded category of forces. These are commonly categorized by strength from the strongest to the weakest as ionic forces, hydrogen bond forces, dipole to dipole forces and van

der Waals forces. Molecular modelers lump ionic, hydrogen and dipole forces together as Coulomb electrostatic and maintain a separate category for van der Waals. For a much more detailed treatment of the theory of intermolecular forces the reader is referred to an excellent text dedicated to the subject "The Theory of Intermolecular Forces". (Stone, 2000) Non-bonded electrostatic forces are obvious candidates at first because they can be very strong, they exist and operate between each and every charge in the system (maintaining conservation of energy) and they do extend to infinity. However, referring back to our categorization of forces, we are reminded that we are referring here only to electrostatic forces between atoms not actually bonded together (covalently) implying minimally large separation distances. Since electrostatic forces diminish exponentially as separation distance increases it becomes unlikely that electrostatic attraction alone would be large enough to move a massive DNA molecule let alone the additional masses associated with chromatin structure.

From here we can refer to an analytical study of long range intermolecular interactions between CG-CG nucleotide pairs and TA-TA nucleotide pairs facing each other across a double strand break of DNA. (Pinchuk & Vysotskii, 2001) Although the intent of the analysis was a better understanding of processes by which double stranded DNA breaks repair themselves many of the author's calculations are pertinent to this investigation. The authors calculated the total energy between ends of a break and graphically presented the results versus distance in angstroms. In addition to the fact that the results were not monotone (the forces were attractive *and* repulsive) the total energy for distances beyond 10 angstroms never exceeded 0.06 eV including electrostatic interactions from fractionally charged phosphate groups. Although they actually extend to infinity, in practice electrostatic forces are sometimes even considered

negligible beyond distances of 1nm (10 Angstroms) and are cutoff as a practical tradeoff in cpu time for some MD simulations. (Darden, Perera, Li, & Pdersen, 1999; Guvench & Alexander D. MacKerell, 2008)

In addition to the relatively large distance spans involved with non-bonded interactions there are geometric variables as well. Considering that the HP phenomenon is *sequence* based and a double stranded helix segment of DNA contains the sequence information *within* the molecule, the charged phosphate groups that comprise the backbone take on the form of a geometric barrier. It is unlikely that non-bonded forces from nucleotides alone could operate with sufficient attractive strength through the atomic skeleton of the backbone to account for the observed phenomenon. Consider that the exposed outer backbone of the molecule is comprised of highly electrically charged phosphate and sugar molecules that physically interfere with the exposure of the inner nucleotides to the outer environment. This implies that the base pair sequences from physically separated molecules can't clearly 'see' each other electrically or proximally. Obviously there are deep theoretical implications we are not considering here so we shall simply conclude that at a minimum the backbone of the helix affects the dielectric constant of permittivity between nucleotides of different DNA molecules. Non-bonded forces certainly may be synergistic with the primary mechanism but alone are insufficient to explain it completely. The concept of backbone interference is illustrated by the figures below. Figure 8 is a simple 20 base pair molecule displayed using the common CPK or "ball and stick" graphical representation. Figure 9 is the same molecule represented by spheres corresponding to van der Waals force effects or VDW. Figure 10 shows the molecule tilted slightly demonstrating how

charges on the backbone shield and dominate the inner structure vastly complicating the dielectric effect between nucleotides.

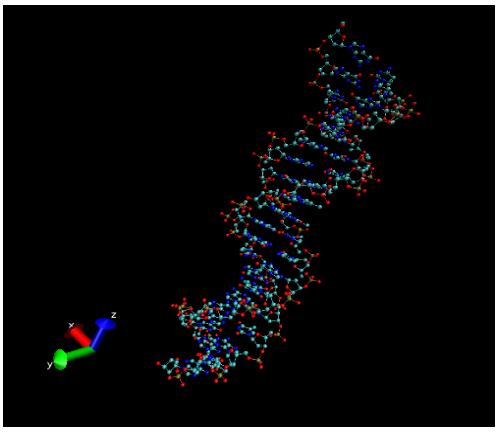


Figure 8: Ball and Stick Graphical Representation

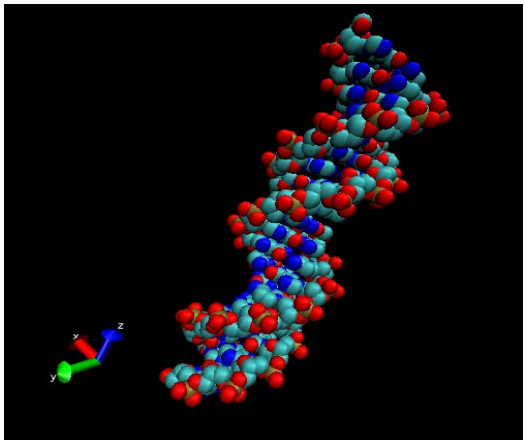


Figure 9: VDW Graphical Representation

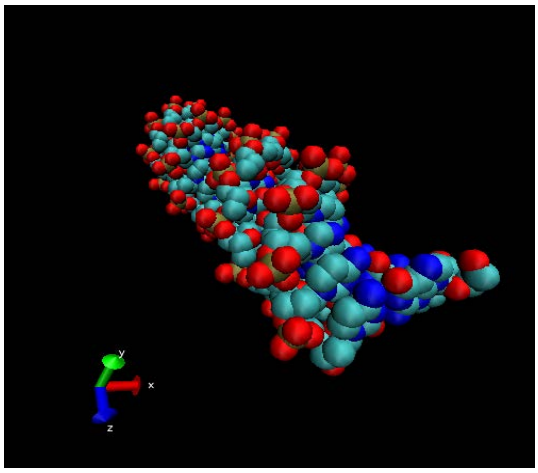


Figure 10: Interference Effect of Backbone Charges

Is Homologous Pairing Based on Diffusion?

In a 1997 study of inter phase chromatin movement it was suggested that chromosome movement like meiotic bouquet formation are the results of diffusion alone. (Marshall et al., 1997) A caveat was added that "a given chromatin segment is confined to a sub region of the nucleus" implying the existence of a highly defined nuclear architecture in order to explain the obvious constraint on the observed movement. Strictly speaking, diffusion is commonly defined as the movement of a substance from an area of high concentration to an area of lower concentration driven by the energy of Brownian motion. If the substance diffusing is considered to be chromatin than the concept of 'concentration' loses its meaning especially since DNA concentrations actually *increase* during bouquet formation and HP. The Brownian motion component of the suggestion does make sense when considered as the energy supply for the movement. The remaining catch is that by definition Brownian motion is a zero-mean stochastic

process (Mathworks, 2009). This is obviously not what we observe. As was determined by the (Carlton et al., 2003) simulation experiment a directed bias is required for these processes to occur.

Why is this Important?

The implication is that there must be an unknown intrinsic property (Baldwin et al., 2008) of a double-helix DNA molecule that directionally biases Brownian motion and that is based solely on base pair sequence. The significance of this characteristic cannot be understated. If this property could be identified and characterized it could provide a common denominator unifying the currently disjointed fields of biology with physics and chemistry. Perhaps even more significant are the possible applications of such a property. The most obvious application of course is the field of drug design. Researchers might design molecules or nano-devices that could identify and home in on DNA molecules based on the nucleotide sequence. Targeted treatments for genetic diseases could ultimately be realized.

Notwithstanding, the Imperial College experiment has set the stage for an exciting new direction of research and most likely new paradigms in chemistry, physics and biology. This is where my research interest lies. As will be discussed in Chapter 2, I believe that currently the most promising method available to further explore and characterize the underlying mechanism of homologous pairing is molecular dynamic simulation. In light of the Imperial College experimental results I believe that one can safely assume that homologous pairing behavior is not dependant on the enormous chromosome scale. Rather this research hypothesizes that

homologous pairing behavior is more strongly correlated to the much smaller scale of a 293 base pair DNA segment. This research endeavors to investigate this hypothesis within the purview of current simulation capability.

CHAPTER 2: PROGRESS OF MOLECULAR MODELING

Chapter 2 Abstract:

In the previous chapter the historical record of research on the DNA molecule was reviewed. The process of homologous pairing was identified as a potentially high value research topic that exists within the interface of biology, chemistry and physics. The chapter alluded to the benefits of further investigating HP with computational simulation, specifically MD simulation. With that as an objective chapter 2 reviews the historical progress of molecular modeling. It begins with the earliest mid 19th century attempts to model atoms and molecules with drawings and progresses to the first application of a computer to the problem in 1957. Rapid progress is described up through the late 1990s where the chapter identifies molecular mechanic (MM) models as an appropriate tool for this investigation and focuses on current research in that particular area. After quantum mechanical (QM) models are differentiated from Newtonian MM models, MM models are affirmed as the most appropriate choice for this research mainly because of current hardware capabilities and the potential system sizes that need to be simulated. The theory underlying MM models is explored in detail and concepts pertinent to our objective are addressed. The chapter concludes by noting the growing popularity of the MM class of models and their ever increasing fidelity to experimental results.

Historical Progress of Molecular Modeling

Trying to understand the interaction of individual atoms or small groups of atoms on an actual atomic scale is a formidable challenge. The most obvious hurtle to this research is the

sheer size (or lack thereof) of the objects of study. Probably the most ubiquitous tool in the field of biology is the light microscope and most of the existing data has been obtained with that venerable instrument (D. Zickler & Kleckner, 1998). This explains why most of the existing research and the vast majority of HP research consists of data gathered by visual observation. To learn more about HP a new approach is needed.

As mentioned earlier, we now know a molecule is a dynamical system comprised of constantly moving atoms and a wide variety of interacting forces. A valid model of a molecule must therefore contain many multi-variable, inter-related time dependant elements. How do we model such a system? Obviously, modeling a dynamical system of atoms is a significant challenge, especially with pre-computer technology. The first attempts at doing this began between 1858 and 1861 when written drawings of carbon chains with lines drawn to represent bonds between atoms and atom groups were used to represent early molecular formulas. Archibald Couper, Friedrich von Stradonitz, and Aleksandr Butlerov independently introduced the general rules of valence for organic chemistry and the term "chemical structure" was first used to describe these molecular formulas. The first known physical model of a molecule appeared in 1865 when August Wilhelm Hofmann used croquet balls joined by sticks to describe several carbon compounds in his lecture "On the Combining Power of Atoms". During the late 1920's and early 1930's the first mathematical models of molecules were developed. These models, later known as "force field" models, were developed by spectroscopists whose objectives were to reproduce and predict vibrational frequencies. A "force field" was a model that considered individual forces between every atom of a molecule. These early models used the quadratic form of Hooke's Law to approximate the potential energy between atom pairs as if

they were connected by springs. Little attention was paid to these models until 1946 when the idea of incorporating Newtonian mechanical variables for bond-stretching, angle bending and torsional vibrations came about. The resulting empirical force fields represented the introduction of what is now known as the molecular mechanics method of calculating molecular structures.

Even though force fields continued to improve, the chemical community at large took little notice of the work until 1953 when the first Monte Carlo simulations were performed with very simple models of molecules that used spheres and discs. The work was titled "Equations of State Calculations by Fast Computing Machines" and introduced what is now referred to as the Metropolis Monte Carlo algorithm for simulating movement of molecules at an atomic scale. (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953) This method demonstrated great potential for the application of computers to molecular studies and is still appropriately used today in specific applications. However, if one is interested in the actual dynamics of a system Monte Carlo analysis is not helpful.

As mentioned earlier, 1953 was also the year that James Watson and Francis Crick, building on the work of many other contemporary scientists like Maurice Wilkins and Rosalind Franklin, brought molecular modeling out of the scientific community and into a world-wide discussion when they presented their model of the structure of DNA. The Watson and Crick "ball and stick" model is arguably the most famous molecular model in the world with modern variants known as CPK models still being widely used in classrooms today. In spite of its popularity this "skeletal" model is static in nature and imprecise in dimension. Ball and stick models provide insight to the 3 dimensional geometry of a molecule but show little else.

Then in 1957 two theoretical physicists Adler and Wainwright outlined a method to calculate exactly the behavior of several hundred interacting particles by applying the classical laws of Newtonian mechanics. Their study was published in the Journal of Chemistry and Physics and was titled "Phase Transition for a hard sphere system." This study was the first application of computational simulation to molecular dynamics and as such is the ancestor of modern MD simulation. (Adler & Wainwright, 1957) This was a huge computational burden for computers of the late 1950's so results were limited. In 1958 Andre Dreiding invented the "Dreiding Stereo Model". This was a highly accurate (and expensive) model made up of modular elements carefully designed to account for the correct number of bonds and specific angles for the particular atom being modeled. These elements allowed the modeler to build up very precise 3 dimensional models of a crystal structure with dimensions carefully scaled up from the true nano-meter distances of the atoms being modeled. In 1961 the first known paper detailing the use of a computer doing molecular dynamic calculations on a molecular structure using a force field model was published in the Journal of the American Chemical Society. (Hendrickson, 1961) In 1965 the "steepest descent algorithm" was introduced by Wiberg as a method to optimize structure geometry and assist in conformational analysis. (Schlecht, 1997; Wiberg, 1965) Progress in the field was commensurate with computer development right through the late 1970's as major force field formulations began to mature. Very sophisticated simulations began to appear in the literature including complex proteins, oligosaccharides, carbohydrates and even polypropylene. (Schlecht, 1997) The 1980's ushered in the age of the personal computer and the graphical user interface (GUI) which brought computational molecular modeling within easy reach of the average scientist. Model development continued

rapidly right along with ever increasing computer power and keen interest from a growing community of modelers. The 1990's saw the astounding growth of the internet and all things graphical, including graphical capabilities of molecular modeling software. Brilliant graphical displays of entities that previously only existed in our imaginations have fundamentally enhanced the value of such simulations. The rate of change in the field of molecular modeling over the last 20 years makes it almost impossible to identify high value landmark research until more time passes and the discipline matures further.

Fortunately for the purpose of this investigation, the molecular mechanic class of models has begun to converge. With the exception of a new class of QM-MM (quantum-molecular mechanical) hybrid models under development, the theory and principles underlying the current generation models is now widely accepted and showing better and better agreement with experimental result, as will be discussed below.

Modern Molecular Force Fields and Simulations

Molecular modeling today can be loosely classified into 2 general categories differentiated by their fundamental approach to describing molecular systems. The first approach and the most extensively developed is based on quantum mechanics (QM). Quantum mechanics, in its most basic sense, is the science of matter and energy at the atomic level. It follows naturally then to use QM methods to model systems of atoms. Quantum mechanics accurately describes *sub-atomic* particles down to the electron level. With regard to modeling an atomic system two of the most important things that QM can describe are the spin of a particle and the discreteness of energy. If a modeler is interested in any system property based on

electronic distribution within an atom QM methods must be used. *Ab initio* (from first principles) QM methods are again preferred when little or no experimental data exists and the system under study is very small. QM methods are also required when trying to model more than just energy and geometric conformational behavior. The quantum mechanical wave function can predict *ab initio* any molecular properties including covalent bond breakage and formation that agree very closely with experimental results. (Leach, 2001)

Despite the obvious appropriateness of using QM methods for molecular modeling there is a drawback. A significant challenge when using a QM approach is the difficulty in obtaining an acceptable wave function characterizing the motion of sub-atomic particles from which to calculate the energy of the system. In 1925 Erwin Schrödinger successfully developed an equation that accurately describes the *evolution* of the wave functions needed for QM models. What is now commonly referred to as "The Schrödinger equation" is a second order partial differential Eigen value equation. The fully time dependant form of the equation is frequently written as follows:

$$\left\{-\frac{\hbar^2}{2m}\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}\right)+V(r,t)\right\}\psi(r,t)=i\hbar\frac{\partial\psi(r,t)}{\partial t} \qquad \textbf{(1)}$$

Applying this equation to a sub-atomic particle like an electron the variables would be defined so that m would equal the mass of the electron, r is a $\vec{r}(x,y,z)$ vector representing position in Cartesian coordinates and $V(r,t)$ represents potential energy acting on the electron (like

electrostatic potential from other electrons or the nuclei) as a function of time t and coordinate position $\vec{r}(x, y, z)$. The variable i is the imaginary number $\sqrt{-1}$, $\hbar$ is Plank's constant divided by 2π and $\psi(r,t)$ is the desired term or wave function. The challenge is great because exact solutions can only be found for systems of 1 or 2 particles. For systems with greater than 2 particles the equation becomes intractable and exact solutions cannot be found. As a result, approximations must be introduced for any system more complicated than a single hydrogen atom. Even though there are currently many excellent computational chemistry packages now available like the GAUSSIAN whose first version was co-authored by John Pople the computational burden is so great even by today's hardware standards that only very small systems limited to approximately 100 atoms or less can be modeled using QM methods. Because of this severe limitation additional review of the QM approach does not further the objective of this investigation but the reader is referred to several excellent texts on the subject (Cramer, 2004; Leach, 2001; Szabo & Ostlund, 1996).

Before discussing the next approach to molecular modeling, molecular mechanics, one very relevant assumption of quantum models needs to be discussed in further detail. As mentioned above the Schrödinger equation cannot be solved for any system involving 3 particles or more so approximations are introduced. An assumption that is key to the foundation of the molecular mechanics approach is illustrated by the Born-Oppenheimer approximation. This approximation is based on the fact that the resting mass of a proton is more than 1800 times the resting mass of an electron. From this large disparity we may assume in most cases that motion of an electron relative to the motion of its nucleus will change instantaneously with a change in

the nucleus. It naturally follows that they can be treated separately. Using this approximation the wave function of the molecule can be written as:

$$\psi_{Total}(r,t) = \psi_{Total}(nuclei, electons) = \psi\ (electrons)\psi\ (nuclei)$$ **(Leach, 2001)** **(2)**

From an energy standpoint this means that total energy can be approximated as the sum of the potential energy of the electrons and the sum of the potential energy of the nucleus.

$$\mathrm{E}_{Total} = \mathrm{E}_{Electrons} + \mathrm{E}_{Nucleus}$$ **(Becker, 2001)** **(3)**

This equation combined with the concept of transferability illustrates the primary concept that allows the second approach to molecular modeling (molecular mechanics) to work at all.(Leach, 2001)

More specifically, the molecular mechanics approach is based on classical or Newtonian physics and does not take into account quantum effects like electronic positions and motions. This approach is characterized by simplifications in its models for molecular systems. Primarily molecular mechanic models are functions of nuclear positions alone, thus the smallest entity considered in the models is an entire atom whereas QM models consider the explicit behavior of nuclei *and* electrons. This simplification is validated by the Born-Oppenheimer approximation and allows the construction of relatively simple models describing only bond stretching, angle bending and angular rotation about bonds. Additionally, as will be illustrated below, the

expanded formulation of Equation 3 and Equation 4 are not chemically accurate because several terms in the formulations are treated harmonically. The use of harmonic terms is considered justifiable when simulating bio-molecules because the simulations are performed at or around room temperature close to equilibrium with no bond breakage or bond formation events. The equations represent a practical balance between chemical accuracy and required simplicity.(Becker, 2001)pg 9. Despite these apparently severe simplifications force field models are capable of producing results that are as accurate as the highest level quantum mechanical calculations while using only a small fraction of corresponding cpu time. (Leach, 2001) pg 165. These models are based on conservation of energy and an empirical energy function. Models based on molecular mechanics are more manageable mathematically for systems of thousands and hundreds of thousands of atoms. Today a model based on a molecular mechanics approach typically consists of a differentiable function of atomic coordinates and a set of parameters describing the energy resulting from intermolecular and geometric interactions that is simply called a "force field".

Within a modern application like a simulation software package, a force field consists of a set of equations that attempt to represent the potential energy function from 2 previously defined molecular properties, bonded and non-bonded interactions. The bonded terms represent covalent bond stretching and compressing, valence angle bending, and torsion potentials generated by rotation around bonds. The non-bonded terms represent Coulomb electrostatics, and a Lennard-Jones approximation of van der Waal's forces. In its simplest form the total energy of the system is represented as:

$$E_{Total} = E_{Bonded} + E_{NonBonded} + E_{Other}$$ **(Becker, 2001)** **(4)**

The E_{Bonded} term is the energy contribution from atoms directly bonded together. The $E_{Non-Bonded}$ term represents the energy contribution of atoms not directly bonded together but close enough to interact electro-statically or by van der Waals forces. Since these are all parameterized models E_{Other} represents a variety of parameters that vary from model to model. Since this is the intended modeling approach for this research we will examine this equation more closely. Expanding the first term of Equation 4 we get:

$$E_{Bonded} = \sum_{Bonds} K_b(b-b_0)^2 + \sum_{Angles} K_\theta(\theta-\theta_0)^2 + \sum_{Dihedrals} K_\chi[1+\cos(n\omega-\sigma)]$$ **(Leach, 2001)(5)**

Thanks to the Born-Oppenheimer approximation it can be safely assumed that the interaction between nuclei will obey Hooke's Law of elasticity for the extension of a spring where the energy is derived from the second degree polynomial $E = \frac{1}{2}kx^2$ where k is the spring constant and x is the distance the spring is compressed or stretched from its steady state position. Applying this to a system of atomic nuclei and summing over all bonded pairs results in the total energy of a system. To illustrate consider Figure 11 in which atoms within the system are divided into subsets of 3 where Atom2 is bonded to Atom1 which is bonded to Atom3.

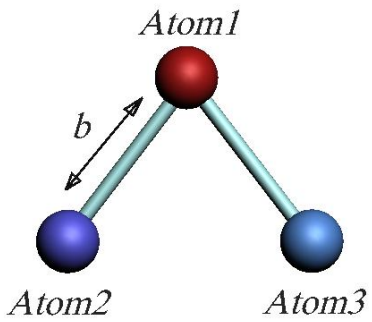


Figure 11: Three Atom subset showing bond stretching

The first term of Equation 4 represents the energy from stretching or compressing chemical bonds and is most illustrative of Hooke's Law. The term is $K_b(b-b_0)^2$ where K_b and b_0 are descriptive parameters for stiffness and steady state or natural bond length and b is the inter-atomic distance between pairs (usually in Angstroms). The second term is derived in exactly the same way. In this case Hooke's Law is applied to the bending of the angle θ formed between bond vectors $\overrightarrow{A1A2}$ and $\overrightarrow{A1A3}$ as in Figure 12.

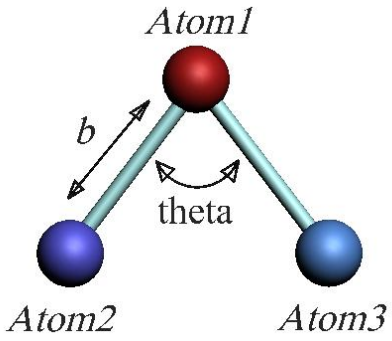


Figure 12: Three Atom Subset shows Angle Bending

K_0 is a parameter that describes the bending stiffness and θ_0 (theta zero) describes the steady state or natural bond angle while $\theta-\theta_0$ is the deflection amount resulting from the bending

force. This results in the second quadratic term $K_\theta(\theta-\theta_0)^2$. The final term is obtained by describing the 'twisting' of dihedral angles around other bonds. See figures below;

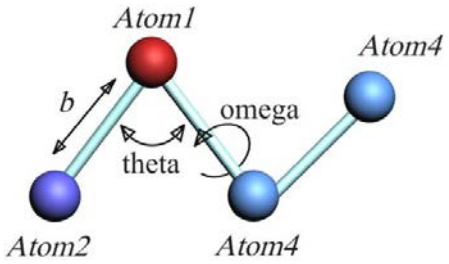


Figure 13: Four Atom Subset showing Dihedral Twisting

Unlike bond stretching and angle bending the potential energy of torsional force around a bond is not linear but varies as it proceeds through 360 deg of rotation and so needs to be expressed by a sinusoidal function. A cosine series expansion is most often applied resulting in $K_\omega[1+\cos(n\omega-\sigma)]$ where the periodicity or number of cycles per rotation around the bond, is n and the phase is σ while K_ω is the familiar force constant.

Moving on to non-bonded energy components we expand the second term of Equation 4 to get:

$$\mathrm{E}_{NonBonded} = \sum_{\substack{NonBonded\\ pairs,ij}} \left(\varepsilon_{ij} \left[\left(\frac{R_{\min,ij}}{r_{ij}} \right)^{12} - 2*\left(\frac{R_{\min,ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{r_{ij}} \right) \quad \textbf{(Leach, 2001)} \qquad \textbf{(6)}$$

This term can be further broken into 2 parts. The $\varepsilon_{ij}\left[\left(\frac{R_{\min,ij}}{r_{ij}}\right)^{12}-2*\left(\frac{R_{\min,ij}}{r_{ij}}\right)^{6}\right]$ component is a widely accepted approximation of attractive dispersion forces and repulsive Pauli exclusion forces collectively known as the Lennard-Jones equation.(Guvench & Alexander D. MacKerell, 2008) Within the model it is usually referred to as the 'van der Waals' term. The operation of the term is quite simple. As 2 atoms $Atom_i$ and $Atom_j$ not directly bonded to each other move closer together they experience an attractive force that increases with decreasing distance r_{ij}. These atoms would accelerate toward each other and eventually collide if not for the repulsive Pauli exclusion force that begins to operate at small values of r_{ij}. At the point of minimum energy for the bond the distance between the 2 atoms achieves $R_{\min}$ meaning the attractive force exactly equals the repulsive force. As the atoms continue to move closer together beyond the energy minimum the repulsive force increases according to a much steeper exponential than the attractive force. This results in quickly increasing repulsive energy pushing the atoms apart. The repulsive force is modeled by $\left(\frac{1}{r}\right)^{12}$ where r is the inter-atomic distance. For computational efficiency the repulsive force term was simply assigned to be the square of the attractive force term $\left(\frac{1}{r}\right)^{6}$. Although not exact this has proven to be a very adequate model of the actual repulsive Pauli function within MD simulations (Guvench & Alexander D. MacKerell, 2008). The ε_{ij} term is an empirical parameter of the model that is based on the type of atoms

interacting. The $R_{\min}$ term is the minimum energy distance parameter that again is a function of the type of interacting atoms. The second component $\frac{q_i q_j}{r_{ij}}$ models electrostatic forces between atom pairs and is simply Coulomb's Law. The r_{ij} term is again the inter-atomic distance and q_i and q_j represent the effective charges on $Atom_i$ and $Atom_j$ respectively.

To be thorough, it should be pointed out that a typical MD force field is conspicuously missing an explicit term for hydrogen bonds. In all the force fields reviewed for this application none contained explicit terms for hydrogen bonds. The general consensus currently is that hydrogen bonds are handled adequately by the above mentioned terms alone without having to be modeled separately. Only in certain cases where the angular dependence of hydrogen bond energy causes a significant variation from QM results does this become a problem. (Morozov, Kortemme, Tsemekhman, & Baker, 2004)

One last characteristic of MD force fields that is relevant to this investigation is the parameter optimization process. As mentioned earlier the E_{Other} term represents a variety of parameters that vary from model to model. During the optimization process the output of a particular force field is rigorously compared to unique sets of target data including spectroscopic, crystallographic and thermodynamic data as well as *ab initio* results computed from quantum mechanical models. The most common experimental target data include heat of vaporization, density, X-ray diffraction structures, vibrational spectra and conformational energies. Good examples of quantum mechanically computed target data are dipole moments, conformational energies, energy minimum geometries and vibrational spectra. (Guvench & Alexander D.

MacKerell, 2008) After comparison some or all of the force field parameters are adjusted to force the model outputs to match the target data. This lengthy iterative process results in excellent agreement between model output and target data. The assumption of transferability allows the application of models highly optimized for a small number of atoms to then be applied to the study of a much wider range of systems. (Leach, 2001)

With all of the needed forces defined the partial derivatives with respect to Cartesian coordinates of those forces are plugged into Newton's equation of motion $F = ma$ and summed over the entire system. This process is repeated iteratively simulating the entire systems behavior through time on the basis of classical mechanics. The results of each simulation step can be stored and later assembled into a series of "snapshots" depicting atomic locations. This assembly of snapshots is known as a "trajectory" and is used extensively for applications ranging from simple visualization to energy surface analysis.

As a final note to this review of molecular modeling research it should be noted that recent advances in computer hardware and freely available molecular simulation software has swept computational research into the mainstream. The choice of models and simulation packages is growing dramatically and published simulations involving hundreds of thousands of atoms over micro-seconds of time are becoming routine.

CHAPTER 3: RESEARCH METHODOLOGY

Chapter 3 Abstract:

Chapter 2 concluded by noting the growing choice and availability of molecular modeling software and high performance low cost hardware that can be utilized for this research. Chapter 3 begins with a summary observation of the large research gap associated with HP and presents 4 hypotheses designed to examine the phenomenon from a new perspective. The first hypothesis investigates the question, can current generation molecular dynamics models produce closure as representative of DNA segregation observed in the Imperial College laboratory experiment? The second and third hypothesis investigates the questions, do current generation molecular dynamics models produce resonance between DNA molecule segments and secondly if so, is the resonance harmonized between similarly sequenced molecules in various paired DNA segment configurations? The fourth and fifth hypotheses investigate the questions, do current generation molecular dynamics models produce interacting harmonized resonance between two DNA segments of dissimilar molecular sequence and, finally, does closure between the two DNA segments occur as a result of interacting harmonized resonance? A detailed rationale immediately follows. A research approach, experimental objectives, metrics and mechanisms are defined in detail followed by appropriate experiments capable of testing the hypotheses. The chapter concludes with a discussion of pilot testing and potential pros and cons of a small scale prototype endeavor.

Using MM Models

Given the current availability of well validated models combined with ever-growing computational power it is feasible and prudent to test the foundation of current generation models against a new target data concept derived from HP. Relying on the key concept of transferability mentioned in Chapter 2 I believe that MM force field models in their current manifestations might be capable of reproducing various aspects of homologous recognition. In light of the complete lack of valid theses explaining any part of homologous pairing as well as the results of the Imperial College Study in 2008 that demonstrated conclusively that long distance attraction between similar nucleotide sequences in DNA occurs spontaneously in a protein-free environment (Baldwin et al., 2008), it is reasonable to expect that the underlying mechanism(s) of homologous pairing, if not HP itself, should be a part of validation target data for models. An hypothesis is presented below that explains how HP behavior might already be found in current molecular models.

It is now apparent that all mechanical bridging and protein chain based models of HP do not provide a sufficient or adequate explanation of the phenomenon. The only remaining *known* mechanisms are the 2 non-bonded atomic interactions and they also are not sufficient to adequately explain HP. A new investigative approach is needed along with new hypotheses.

As mentioned in the opening statement for this chapter I believe the most logical starting point to further investigate HP is to assume that DNA strand segregation is a small scale example of the much larger phenomenon of homologous *chromosome* pairing. Small DNA strands (as opposed to entire chromosomes) can easily be modeled with current generation molecular

models. The following five questions underlie the five main hypotheses of this research. They are:

- Are current MD simulators able to reproduce the segment closure observed in the Imperial College experiment?
- Are current MD simulators able to produce resonance between identically sequenced DNA segments?
- If so, is the resonance harmonized between the identically sequenced DNA segments?
- Do DNA molecules with dissimilar nucleotide sequence mechanically interact to create interacting harmonized resonance?
- Does interacting harmonized resonance cause two DNA segments to move closer together resulting in the observed closure?

The following formal hypotheses are a first step in this approach.

Research Hypotheses

Hypothesis (Simulate Observed Closure):

H$_{\text{(Simulate Observed Closure NULL)}}$ **: =** Current MD force field models when used to model DNA molecule segments, contain sufficient variable terms and parameters to describe and reproduce directed segregating movement (closure of the segments) as previously observed by the Imperial College team between two Phi X 174 DNA molecules. The alternative is therefore

H$_{\text{(Simulate Observed Closure ALT)}}$ **: =** Current MD force field models when used to model DNA molecule segments, DO NOT contain sufficient variable terms and parameters to describe and

reproduce directed segregating movement (closure of the segments) as previously observed by the Imperial College team between two Phi X 174 DNA molecules.

Rationale:

Given the complexity of molecular mechanical models combined with the various simplifications over the real-world systems they represent it is difficult to predict the outcome of these tests. Since this is an unprecedented application we do not know if current models are capable of reproducing any of these phenomena, theorized or observed. Considering only the growing body of published results showing agreement with a wide variety of experimental data it is reasonable to expect a molecular mechanical model will *at least* reproduce homologous segregation behavior. Further speculation is unwarranted until this test has been performed.

Hypothesis (Resonance):

H(Resonance NULL) : = Current MD force field models when used to model DNA molecule segments in a condensed phased solvent contain sufficient variable terms and parameters to reproduce theorized molecular resonation in the form of frequency content found in water between the segments.

H(Resonance ALT) : = Current MD force field models when used to model DNA molecule segments in a condensed phased solvent DO NOT contain sufficient variable terms and parameters to reproduce theorized molecular resonation in the form of frequency content found in water between the segments.

Hypothesis (Harmonized Resonance):

H$_{(\text{Harmonized Resonance NULL})}$ **: =** Current MD force field models of DNA molecule segments in a condensed phased solvent produce theorized molecular resonation in the form of frequency content above and beyond the expected normal frequency levels found in water between the segments. The alternative being:

H$_{(\text{Harmonized Resonance ALT})}$ **: =** Current MD force field models of DNA molecule segments in a condensed phased solvent DO NOT produce theorized molecular resonation in the form of frequency content above and beyond the expected normal frequency levels found in water between the segments.

Rationale:

In the absence of contradictory evidence it is reasonable to assume that MD models can reproduce a theorized phenomenon of vibration in the solvent between DNA molecule segments as well as Harmonized molecular vibration (resonance). MD calculations use Hooke's Law to individually describe a large number of independent oscillations of each molecule within a system. These oscillations are commonly referred to as 'normal modes' and are *generally* expected to occur in the range between 10E12 and 10E14 Hz. For N atoms in a particular molecule the number of normal modes will be either 3*N-5 or 3*N-6 depending on the linearity of the molecule. For example a single water molecule consists of 3 atoms in a bent configuration (not linear). Using this rule of thumb it would vibrate at 3*3-6=3 different frequencies or 3 normal vibrational modes. The cumulative effects (if any) of many of these discreet vibrations within a small system of solvent is currently unknown. If these individual vibrations somehow

become integral or additive with adjacent molecules they would likely manifest a quantifiable *periodic* variation in the system pressure. The relative distribution of frequency and magnitude of such a variation is also unknown. Lastly, with one or more DNA molecules added to the solvent the effects (if any) on inter-system pressures are again, unknown. If some unique property of DNA molecular structure causes those independent oscillations to converge into lock-step within a DNA segment it is reasonable to assume that an MD simulator could reproduce the activity. It is feasible that this superposition of DNA specific normal modes could result in relatively large magnitude inter-molecular pressure variations at lower fundamental frequencies within close proximity to a DNA segment

Hypothesis (Interacting Harmonized Resonance):

$H_{(\text{Interacting Harmonized Resonance NULL})}$ **:** = Condensed phase solvent immersed DNA molecule segments with similar nucleotide sequences mechanically interact resulting in harmonized resonance between separate DNA molecules. The simple alternative is therefore:

$\mathbf{H}_{(\text{Interacting Harmonized Resonance ALT})}$ **:** = Condensed phase solvent immersed DNA molecules with similar nucleotide sequences do not mechanically interact through harmonized vibrations.

Rationale:

This hypothesis is a simple expansion of the previous hypothesis to include interaction with another DNA molecule. The unique structure of the DNA double helix might broadcast by vibration resonance a frequency and magnitude that are dependant mainly on the nucleotide base pair sequence. This interacting harmonized resonance might exhibit a magnitude far greater than

typical normal mode vibrations most likely at a lower frequency. The intrinsic structural characteristic of each helix strand may determine the frequency, magnitude, and other harmonics similar to how organ pipes determine notes in an organ. This intrinsic property of the structure might cause normally random thermal vibration to synchronize within the hydrophobic region of the double helix causing segments of the DNA to emit longitudinal pressure waves out into the surrounding water environment. These waves might simply be transverse compressions and rarefactions of the water molecule bonds surrounding the double helix. If a second DNA molecule of similar sequence within close physical proximity were to be exposed to these waves moving through the solvent interaction at a higher level might occur causing further concentration or superposition of vibrational energy into fewer and lower frequencies. This concentration of energy could effectively amplify certain pressure variation frequencies at the expense of others.

Hypothesis (Resonance Causes Closure):

H$_{\text{(Resonance Causes Closure NULL)}}$ **:** = Interacting harmonized resonation produces an aggregate force between the 2 macro-molecule segments resulting in simulation of the same directed motion and segment closure as observed by the Imperial College team between two Phi X 174 DNA molecules . The straightforward alternative is therefore:

H $_{\text{(Resonance Causes Closure ALT)}}$ **:** = The inter-molecular closure observed during the Imperial College experiment is not the result of interacting vibrational resonance but rather a completely different mechanism.

Rationale:

It is likely that if the period and magnitude of the waves postulated in Hypothesis *(Interacting Harmonized Resonance)* are a direct function of the nucleotide sequence, then the second interacting molecule consisting of exactly the same or similar nucleotide sequence located within relatively close physical proximity might react in more ways than just resonating. Perhaps, given certain physical alignment conditions, resonation and/or amplification of these waves between the 2 molecules might occur. This concentration of energy into specific frequencies might produce either of 2 possible conditions resulting in directed motion, either a saw tooth potential force between the DNA molecules or an asymmetrical boundary on the water molecules between the DNA. A saw tooth force function could result in a Brownian ratchet behavior while asymmetrical boundary conditions will produce a non-zero flux resulting in a directional bias on the Brownian motion. Either condition would cause directed movement.

Research Approach

These 5 hypotheses are heavily inter-related and rely upon one another. The first step in testing these hypotheses and perhaps identifying the mechanism of interest must begin by performing additional validation of a current force field model using DNA segregation as the target data paradigm. The following simulation investigation will accomplish a rudimentary validation and generate critical data necessary for further insight into harmonic resonance as a possible mechanism for the phenomenon.

Experimental Objective and Variables

Since the essence of all 5 hypotheses lies in two quantifiable variables, vibrational resonation and closure, a 2 pronged parallel investigation is suggested. Each prong is a computational experiment utilizing a molecular dynamic simulation. Before considering the details of each experiment suitable metrics must be defined to help in determining experimental parameters as well as analyzing the outcome. Quantifiable metrics are needed for both the independent input variables as well as the dependant output variables. The 2 dependant variables will be closure and frequency content. The 2 independent variables will be sequence similarity and geometric location of the center of mass of DNA segments.

Closure

A good measure of movement for an entire molecule is to consider the movement of the center of mass of that molecule relative to its environment or relative to a second molecule. For this investigation it will be defined as movement of the center of mass of a molecule in a 4 molecule system relative to the center of mass of opposing molecules rather than the environment. The result will be a scalar quantity in units of angstroms that may change over time. The null hypothesis will hold if significant negative values indicating closure between molecules in the system has occurred. Positive values would indicate separation has occurred. The alternate hypothesis will be true for values greater than or statistically equal to zero.

Frequency Content

A good measure of frequency content is the unit-less measure of signal to noise ratio of the frequency spectrum of a data set. The Fourier Transform of system pressures will be examined for the existence of frequency components. Any component with a signal to noise ratio that is statistically significant will be considered evidence of resonation. The hypothesis will hold true for any value that is statistically significant. The null will hold if there are no statistically significant outcomes.

Sequence Similarity

The qualitative degree of sequence similarity is comprised of many variables. The most obvious are length of sequence, number of exact positional matches, length and number of contiguous matches and phase difference between contiguous matches. Comparative ratio's of base pair type percentages of the whole are likely to be informative as well. Each of these variables should represent a proportionally weighted term of a similarity function that should be used to quantify the degree of sequence similarity. For future research this will be entirely appropriate but to avoid un-necessarily complicating the analysis portion of this investigation a simple binary approach will be used. The sequences will either be similar or dissimilar regardless of length. No other variables shall be considered.

Geometric position

In the complete absence of precedence the choice of geometric position for each molecule is purely arbitrary. In order to establish a starting point a single assumption will be made regarding a DNA molecule, the effect of sequence should be cumulative along the length of the molecule. If this assumption holds then 2 things can be extrapolated as a result, the effect will be more pronounced at the lengthwise ends of the molecule and the effect will be more pronounced the longer the sequence. This assumption can be easily tested with 4 simple configurations, parallel, skew, end-to-end and perpendicular "T". The hypothesis will hold if frequency content exists in the output and varies between each configuration. The alternate will hold true if the outcomes are all the same.

An Appropriate Simulator and Force Field

Thanks to an excellent review of current molecular dynamic force fields entitled "Comparison of Protein Force Fields for Molecular Dynamics Simulations" (Guvench & Alexander D. MacKerell, 2008) the choice of force field and simulator becomes much easier. Using the following criteria;

- ✓ They must be low cost or freely available.
- ✓ Force fields must have published results and parameters have been peer reviewed
- ✓ Developed within Academia
- ✓ Because of anticipated computational load application must be scalable to the parallel architecture available to me on STOKES.
- ✓ Must have extensive post simulation analysis capability

- ✓ Because of the specific requirement to capture and analyze a system pressure profile for the experiment associated with the resonance hypothesis, the simulator must have profile output capability

After researching the most popular simulator options available the Nano-scale Molecular Dynamics NAMD (Phillips et al., 2005) simulator developed by the Theoretical and Computational Biophysics Group at the University of Illinois at Urbana-Champaign is the easy choice. It is freely available for download from the official website. It is highly scalable, extremely well documented and free. Among the 4 most popular force fields in the literature today, CHARMM, GROMOS, AMBER, and OPLS-AA, NAMD supports them all although at the moment the user might have to build custom topology and parameter files for OPLS-AA for anything other than proteins. The default force field implementation within NAMD is CHARMM. The simulator is the primary requirement for the research I propose.

Because of the enormous system size and time scales required for this investigation a parallel architecture platform like STOKES here at the University of Central Florida Institute for Simulation and Training is a must. Software for construction and manipulation of molecular systems as well as post simulation analysis is also needed. The Virtual Molecular Dynamics (VMD) modeling software, developed in conjunction with NAMD, is an appropriate choice. This package also is freely available and can be installed on a PC workstation running Windows. A standalone version of MATLAB® is also required for post simulation data analysis and is currently installed and available on the STOKES cluster.

Experimental design to test H(Simulate Observed Closure NULL) and H(Resonance Causes Closure NULL)

Test for Closure

The first experiment is the most computationally intensive and will yield a very basic result. It will consist of a single MD computer simulation of a virtual system of 4 DNA molecules immersed in ionized solvent. Crystal structures are not available for these molecules so they will be manually assembled. The complex choice of base pair sequence is adequately simplified by the binary assumption mentioned above. The choice is further simplified by the opportunity to maintain consistency with the Imperial College experiment. (Baldwin et al., 2008) The obvious decision is to use the same DNA fragments from the Phi X 174 bacteriophage that were used by Baldwin. Two of the molecules will consist of a fragment of the PHI X 174 bacteriophage DNA molecule from base pair 176 to 469. The other 2 molecules will consist of a fragment running from base pair 406 to 699. For computational expediency they will be oriented in a 2 x 2 parallel configuration solvated by a water box extending 10 A beyond the max and min extremities. Remembering that electrostatic and van der Waals forces are frequently considered negligible beyond 10 Angstroms, 10 Angstroms is chosen as a balance between computational burden and a typical cutoff length for non-bonded interactions.(Guvench & Alexander D. MacKerell, 2008) See the figures below.

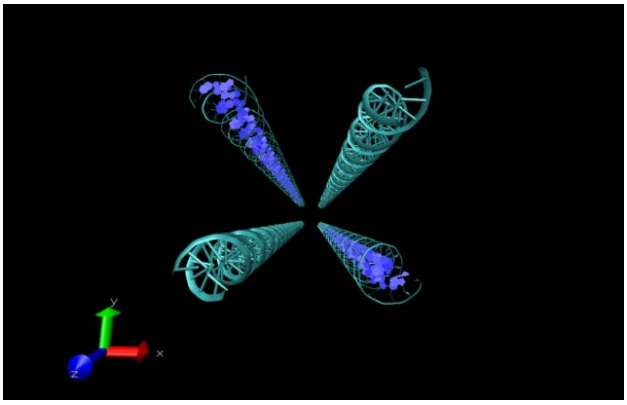


Figure 14: Four DNA Molecule fragments from Phi X174

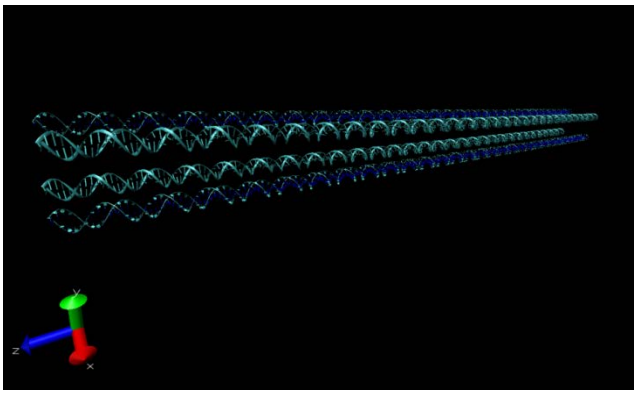


Figure 15: Four DNA Molecule Fragments Side View

This system will be minimized and equilibrated at room temperature consistent with current best practices. Molecular dynamics will be computed for the system in 2 femto second time steps for a total of 1 micro second (or whatever time frame computational resources permit). During the simulation "snapshots" of the molecular coordinates will be saved forming a long run trajectory. Based on a known in-vitro chromosome movement data between 1 and 12nm/second (Carlton et al., 2003) (Gunawardena & Rykowski, 2000) evidence of significant movement (.1 nm) might be observable within 10 milliseconds of simulation time. When the simulation is complete the location of the center of mass for each molecule will be calculated for the first and last frame of

the simulation and the difference calculated. Any resulting difference will be analyzed for significance.

Experimental Design to test H(Harmonized Resonance NULL) and H(Interacting Harmonized Resonance NULL)

Search for Resonance

The second experiment will consist of 10 computer simulations on 5 separate molecular systems. Each system will consist of 2 much smaller DNA molecules again immersed in a solvent box. The sequence for each molecule in 4 of the systems will be 5' TATAAACGCCTATAAACGCC 3' as determined above and match exactly. The 5th system will be intended to highlight sequence effects (if any) and so the sequence of the molecules will be the antithesis of the sequence of all the other molecules. The molecules in each system will exhibit the four previously mentioned geometric orientations; parallel, perpendicular skew, end-to-end and perpendicular "T" all with zero degree axial rotations. The geometries will be obtained by taking the base molecule and applying a transformation matrix to all coordinates accomplishing a 90 deg rotation about the y axis. The transformed molecule will be saved as a new molecule. The base molecule will then be translated in the appropriate X, Y, or Z direction and again saved as a new molecule. The 2 molecules of interest for each system will be loaded together and written out to a single system file. The results will look like the figures below.

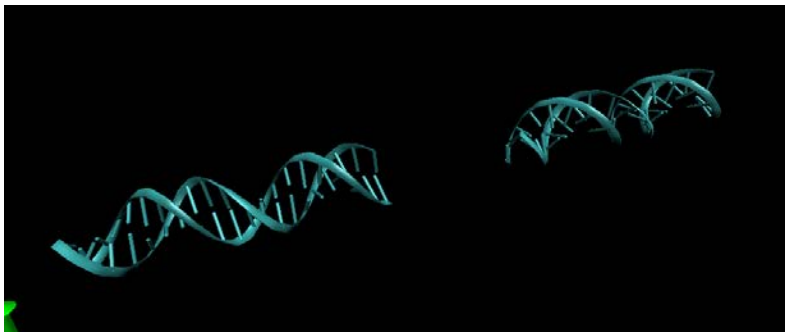


Figure 16: End-to-end, Linear Configuration (Linear for short)

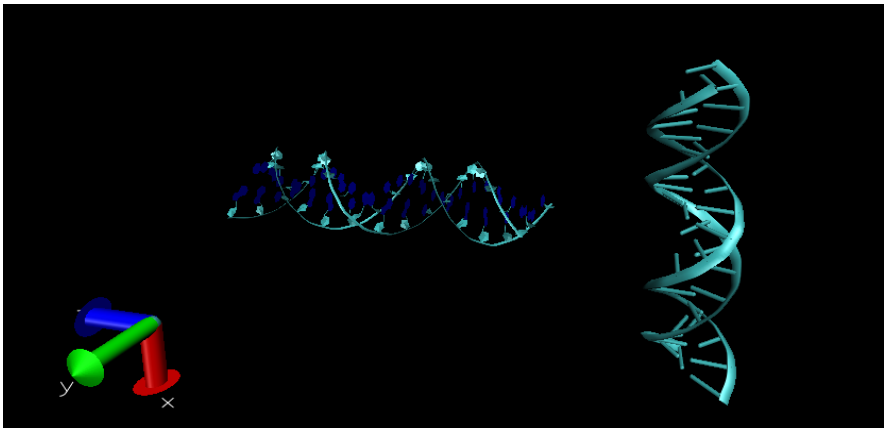


Figure 17: Perpendicular "T" Configuration

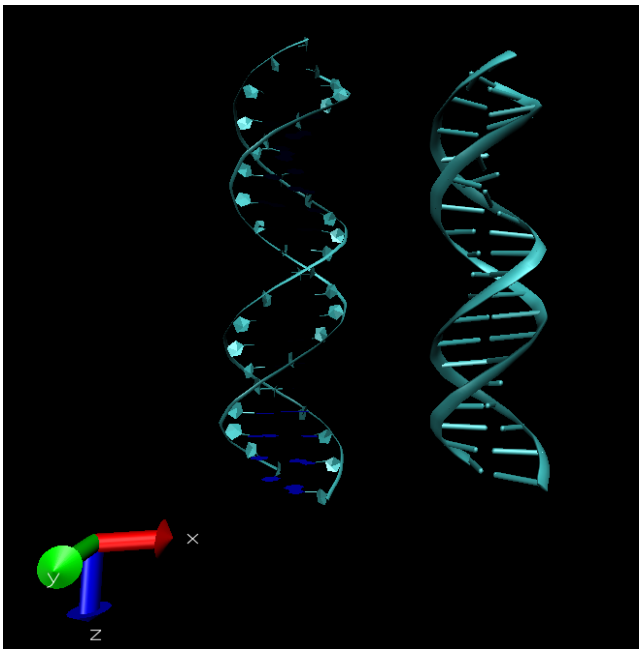


Figure 18: Parallel Configuration

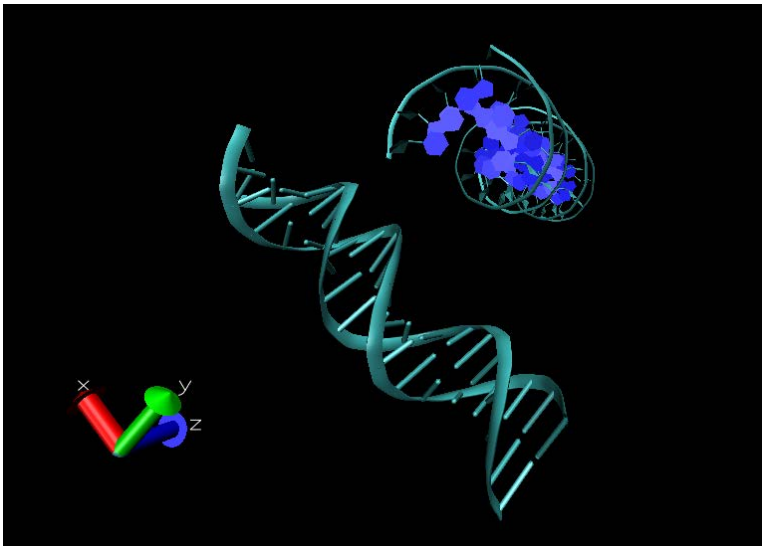


Figure 19: Perpendicular Skew Configuration

The geometry of each system will be oriented in such a way that the gap between the molecules will lie on the XY plane and will be centered on the z axis. Each system will be minimized, heated to room temperature, equilibrated and molecular dynamics will be calculated for a brief

period of time. Since we are looking for a self-starting resonation caused by unknown stochastic molecular events an arbitrary simulation period of 1us (1E-10 seconds) will be used for the first experiment. At the end of a short standard MD run the simulation parameters will be reset to output for every time step a 3x3 pressure tensor profile matrix instead of simple system pressures. The simulation will be restarted and run again for only a short time because of the huge amount of data that will be output. The data output will be parsed offline and the pressure tensor corresponding to the planar slab midway between the 2 molecules will be saved to separate files. Each data set will be input to MATLAB® and a Fourier Transform will be calculated. The results will be examined for frequency content and analyzed for significance. The center of mass for each system will not be examined because the simulation time scales will be too short and the result will likely exhibit only random movement anyway.

Sequence Selection

As mentioned earlier selection of the sequence for the experiment testing $\mathbf{H}_{\text{(Simulate Observed Closure NULL)}}$ and $\mathbf{H}_{\text{(Resonance Causes Closure NULL)}}$ could have been arbitrary but in order to maintain unquestionable consistency with the Imperial College protocol the Phi X 174 bacteriophage base pair sequences from 176 to 469 and 406 to 699 should be used. (Baldwin et al., 2008) Selection of the sequences for the experiment testing $\mathbf{H}_{\text{(Resonance NULL)}}$, $\mathbf{H}_{\text{(Harmonized Resonance NULL)}}$ and $\mathbf{H}_{\text{(Interacting Harmonized Resonance NULL)}}$ again may well be an arbitrary choice but research suggests that particular repeat sequences found in ribosomes may produce better results. (Widlund et al., 1999) The base pair sequence 5' TATAAACGCC 3' monomer is duplicated to create the single repeat sequence of 5' TATAAACGCCTATAAACGCC 3' minimizing the computational burden

while maintaining a single repeat. Without further knowledge of the property of interest it is unclear whether the length to width ratio of the molecule is meaningful.

Experimental Feasibility

The molecular systems proposed in these experiments are enormous and may well have been impossible to simulate with available resources. In order to determine if these experiments could be run at all the basic systems were constructed first to get some idea of actual system size and potential run times. A detailed procedure for constructing these systems is included in the appendix.

For the experiment testing $\mathbf{H}_{(\text{Simulate Observed Closure NULL})}$ and $\mathbf{H}_{(\text{Resonance Causes Closure NULL})}$ a virtual molecule of the Phi X 174 segment running from nucleotide 176 to nucleotide 469 was constructed and solvated with a 5angstrom water box. The resulting system contained 236048 atoms. A second duplicate DNA fragment adds an additional 18638 atoms for a total of 254686 atoms. The system was then re-solvated in a water box and was found to contain more than 500000 atoms. This is pushing the computational limits of much of today's HPC (High Performance Computing) parallel platforms, but the STOKES platform here at UCF can most likely handle this after its recent upgrade.

For the experiment testing $\mathbf{H}_{(\text{Resonance NULL})}$, $\mathbf{H}_{(\text{Harmonized Resonance NULL})}$ and $\mathbf{H}_{(\text{Interacting Harmonized Resonance NULL})}$ each system is much smaller although there are many more iterations of the protocol. The biggest challenges with this experiment will be simulation administration and data management. The total scope of the project is well within current capabilities of STOKES.

Simulator Parameter Selection

Use of the NAMD simulator and the CHARMM force field require a minimum of 120 Simulation parameters to produce a result. All required simulation parameters conform to current best practices and are included in the appendix for the interested reader.

Experimental Predictions/Pilot Testing

The system required to test $\mathbf{H}_{(\text{Simulate Observed Closure NULL})}$ and $\mathbf{H}_{(\text{Resonance Causes Closure NULL})}$ is an enormous construct that will require many hours of preparation and many cpu hours running and managing the simulation after that. It would be prudent to pilot test this scenario before such an investment of time and resources if possible. The 2 primary variables that can be scaled down to achieve a shorter test are size of the virtual system and duration of the simulation. If the length of the DNA molecule is reduced, thus reducing system size, consistency will be lost with the Imperial College result. Also, it is not known whether there is a required minimum chain length for the phenomenon to occur so there is a risk of diminishing or even eliminating a positive result by shortening the chain. Finally, parallel architecture simulators scale best with system size further reducing the general benefit of a smaller system. If simulation run time is reduced the results might indicate a false negative because the inception point of resonation is completely unknown. The chances of an outcome supporting the hypothesis are greatly enhanced by maximizing simulation run times. Furthermore, minimization and equilibration still must be performed and both require lengthy fixed time periods to accomplish regardless of the duration of molecular dynamics. It is unlikely that a pilot test of the PhiX 174 virtual system would be beneficial.

The simulation testing $\mathbf{H}_{(\text{Resonance NULL})}$, $\mathbf{H}_{(\text{Harmonized Resonance NULL})}$ and $\mathbf{H}_{(\text{Interacting Harmonized Resonance NULL})}$ is based on a much smaller molecular system but suggests 5 variations of that system to search for sequence correlation. A feasible pilot test of this system could simply be any one of the proposed geometric iterations, but which one? With so much basic knowledge missing it would be difficult to even guess, underscoring the need for this research as a first step to acquiring that knowledge.

For the Phi X 174 system testing $\mathbf{H}_{(\text{Simulate Observed Closure NULL})}$ and $\mathbf{H}_{(\text{Resonance Causes Closure NULL})}$, in the complete absence of prototype data, the hypotheses predict the DNA molecules within the solvent will migrate closer together and segregate into two groups of 2 corresponding to their sequences.

For the smaller systems testing $\mathbf{H}_{(\text{Resonance NULL})}$, $\mathbf{H}_{(\text{Harmonized Resonance NULL})}$ and $\mathbf{H}_{(\text{Interacting Harmonized Resonance NULL})}$ in the complete absence of prototype data, the hypotheses predict that pressure values taken from the solvent between molecules with the same sequence will show one or more peaks in the frequency spectrum of the data somewhere below the vibrational frequencies of water. Furthermore, the frequency spectrum of pressure variation taken from the solvent between molecules with dissimilar sequences will be relatively flat (i.e. Gaussian).

In the event that neither of the predicted outcomes occurs, a *very significant* previously unknown characteristic of current generation models will have been explored.

CHAPTER 4: RESULTS

Chapter 4: Abstract:

In order to test whether or not current MD simulation will replicate molecular movement and DNA segment closure observed in the Imperial College experiment, a single virtual molecular system, Phi X 174 (the same system used by Imperial College), was constructed and run through molecular dynamic simulation. In what shall be referred to as Experiment #1, simulated molecular dynamics were performed, results recorded, and statistically investigated in terms of the Closure hypotheses presented in Chapter 3. The simulated movement and closure not only did not replicate the Imperial College experiment, but neither statistically significant movement nor closure was observed in the MD simulation. Detailed data and analysis is discussed below. Five more systems were constructed and simulated to investigate $\mathbf{H}_{(\text{Resonance NULL})}$, $\mathbf{H}_{(\text{Harmonized Resonance NULL})}$ and $\mathbf{H}_{(\text{Interacting Harmonized Resonance NULL})}$ that shall be referred to as Experiment #2. At last count taken sometime last year these 2 experiments took more than 12000 cpu hours to complete. *Data and analysis may be found in the Appendix. Discussion is provided in this chapter.*

Highlighting Experiment #1 with the closure hypotheses, the Phi X 174 trajectory files were analyzed and the center of mass was calculated for each helix on the first frame of the trajectory and the very last frame of the trajectory. Using the 3D coordinates for the centers of mass for each molecule the initial distance and the final distance between the center of mass of each molecule was calculated. During 2ns of MD simulation time no-closure was observed. This result was tested for significance with a combination of 2 parametric applications, the t-test and One Way Repeated Measures Analysis of Variance. These tests suggested that the average

positional variations over time of each vibrating molecule relative to its adjacent molecules were not significantly different from zero during the simulation.

Highlighting Experiment #2 with the resonance hypotheses, four molecular systems were initially constructed in accordance with the geometric configurations in Figure 16 through Figure 19. The four systems were carefully constructed to be symmetrical along the z-axis with a similar thickness slab of water between the DNA. The four systems were run through essentially the same routine as Phi X 174 except the special profiling feature of NAMD was utilized to output individual pressures for each system slab. Using this feature all but Ewald sums are computed during the normal simulation run. Because the simulations are run using the Particle Mesh Ewald method (PME) the Ewald contributions had to be calculated with a secondary calculation-only run (no MD was performed) based on the first run trajectories. The Ewald data was then added back into the pressure data from the original run providing the complete pressure picture for each system. (Bhandarkar et al., 2008)

The real time pressure data results of the 4 systems were transformed into the frequency domain with Fast Fourier Transforms and statistically analyzed for spectral content within the water slab between DNA. Large quantities of periodic pressure variation were identified. In view of these findings and the discovery process that led to them the formulation of $\mathbf{H}_{\text{(Interacting Harmonized Resonance NULL)}}$ was revisited. This hypothesis was subsequently refined into a new emergent hypothesis $\mathbf{H}_{\text{(Sequence Relationship NULL).}}$ This refined hypothesis was tested in place of the original. The End-to-End Linear configuration system was re-constructed with a random DNA sequence rather than the energetic sequence in an attempt to correlate DNA sequence to the observed frequency spectrum. The significant frequency content from all 5 systems was

examined for similarities. Although beyond the scope of the proposed research, additional testing was needed because of the refinement of $\mathbf{H}_{(\text{Interacting Harmonized Resonance NULL})}$. A matching program was written to sort the output data from all 5 systems and compile a list of frequency matches. The program was run and the matches found were then grouped by association with the energetic sequence or the random sequence. Of particular note, the grouping revealed 7 specific frequencies that were present in the spectrum when the energetic sequence was present in the molecular system. These seven frequencies were missing from the spectrum when the energetic sequence was missing from the molecular system. Implications of this are discussed.

Alternative Research Method

The proposed research was approved August 19 2009 and began with molecular system building. Sept 1, 2009 a new charging policy was implemented on the UCF STOKES cluster nearly making the research cost prohibitive. An alternative plan was devised using slightly scaled down molecular systems and a smaller custom built 16 cpu cluster temporarily dedicated to this research. All research was successfully conducted with this system.

Experiment #1: Closure Results

The phix174 molecular system was the first to be built. The process started by creating 4 separate DNA molecules, two molecules corresponding to nucleotides176 through 469 and 2 corresponding to nucleotides 406 through 699 thus duplicating the sequences used by Baldwin. The DNA structures were generated using an automated version of the nucleic acid builder

subroutine from the Amber Suite of bio-molecular simulation programs. (Stroud, 2006) All 4 DNA molecules were then loaded into one molecular system in a parallel fashion with an arbitrary separation of 50 Angstroms center to center. While examining the equilibration process for this system it was noticed that because periodic boundary conditions are to be used during the simulation the 50 Angstrom spacing would place the DNA molecules in a "mirrored" position within each periodic replica. A different DNA-DNA spacing was chosen to eliminate this variable. The system was completely re-built with 20 Angstroms of separation providing a spacing that is not an even multiple across replicas. The 4 molecule system was then solvated with water molecules extending 15 angstroms beyond the DNA and the solvent was then ionized to a level similar to that in a natural cellular environment. The system then underwent the 3 required special MD process simulations necessary to place it in a 'natural' state ready for regular MD simulations. The system was first "minimized". This is an abbreviated MD process that allows all the molecules in the system to "relax" to their lowest energy state. This is necessary because when a system is first built (especially systems built from theoretical rather than crystallographic models) there is some minor random variation in exact molecular locations relative to each other creating enormous internal stresses from being placed too close together or too far apart. Minimization is necessary to allow the entire system to gently stabilize to its minimum energy state. The second special MD process simulation the system underwent was "heating". The nature of a newly built molecular system is static. The atoms have no previous velocity information associated with them. It is as if a system is frozen when first built. In order to establish a 'history' of dynamic information to associate with each atom the system must be 'heated' from absolute zero to the desired temperature for simulations to begin (in our case room

temperature). This is accomplished by a short MD simulation where the desired system temperature is adjusted upward in small increments during simulation until the desired temperature is reached. This effectively and somewhat abruptly heats the system up from absolute zero. The third preparatory step is known as “equilibration”. After the system is ‘thawed’ there again exists an un-natural energy landscape within the system similar to that prior to equilibration but not as severe. Equilibration is the process by which the system is allowed to seek a lowest energy state starting point where regular MD calculations can begin. The system is equilibrated until the Root Mean Square Deviation, or RMSD, is low enough to indicate that the system has become stable. The RMSD tells us the amount by which the molecules in the system vary from a particular position in space and is a good indicator of whether or not the DNA in the system is still searching for a lower energy state or not. (Isgro, Phillips, Sotomayor, & Villa, 2007) In the case of Phi X 174 the system showed adequate stability at 3000 minimization steps so the process was continued until 10000 minimization steps were completed allowing ample time. This standard was subsequently applied to all the systems built for this project. After minimization, heating and equilibration the Phi X 174 system was ready for regular MD simulation.

An MD simulation generates large amounts of data sometimes in large individual files so in order to keep simulation and data management easier the simulations were run in sequential segments. Each segment consisted of a varying number steps that was based on practical scheduling of simulation restarts and data management. Throughout the entire process each step represented a 2 femto-second time step for consistency. Each re-start began using the molecular velocities and positions from the last step of the previous segment providing seamless integration

between runs. In order to provide as much simulation time as possible for segregation to occur the simulation was run for 11 consecutive runs. Each run started up where the previous run left off. The first run was started January 1, 2010 and the last run was completed July 10, 2010 accomplishing a total of 1,200,000 steps for a total of 2,400,000 femto-seconds or 2 nano-seconds of simulation time at a cost of 8535 cpu hours. The conformational changes that occurred over the course of the 2 ns are illustrated below.

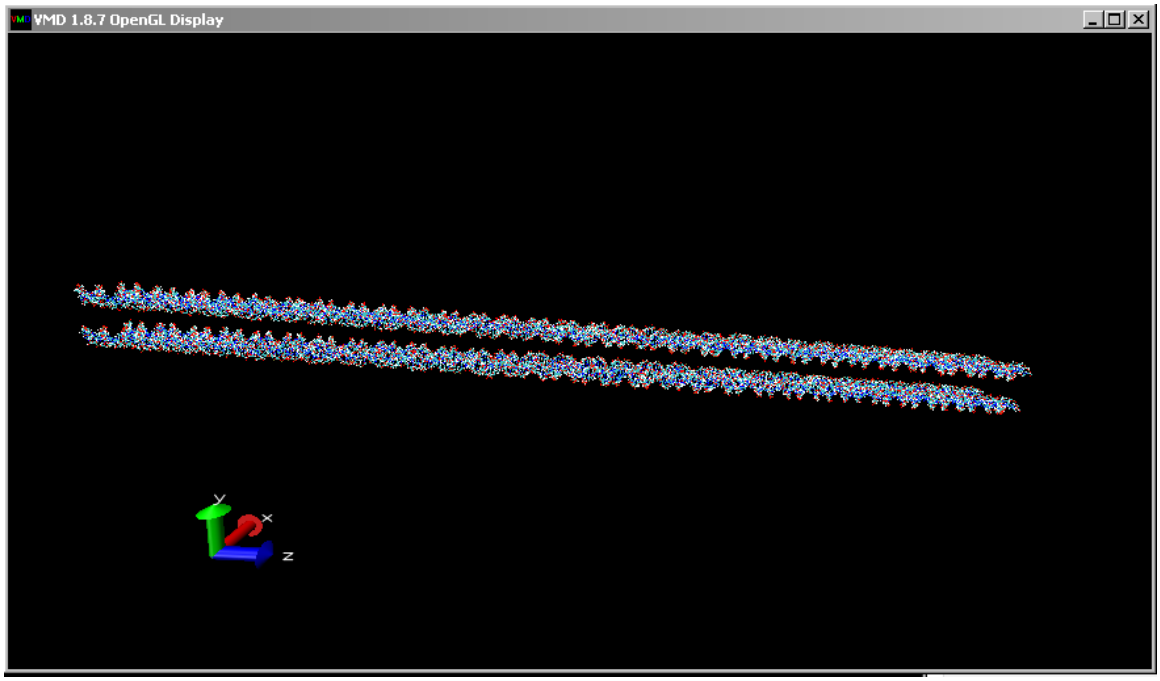


Figure 20: Phix_174 Conformational State before MD Simulation

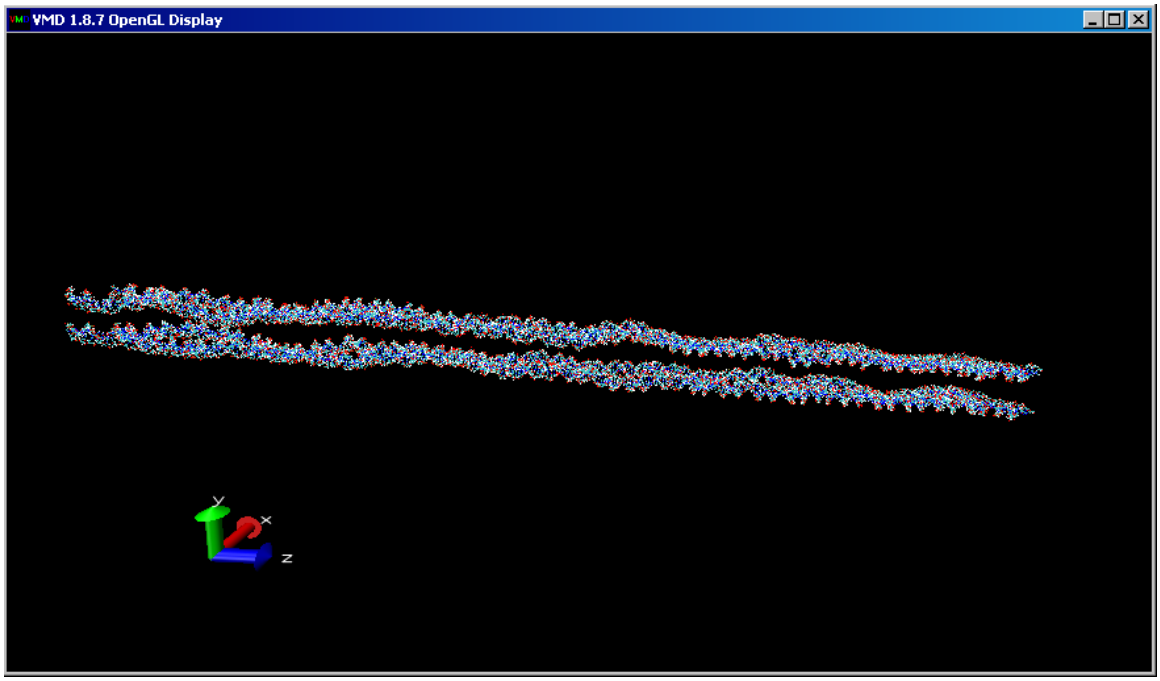


Figure 21: Phix_174 Conformational State after MD Simulation

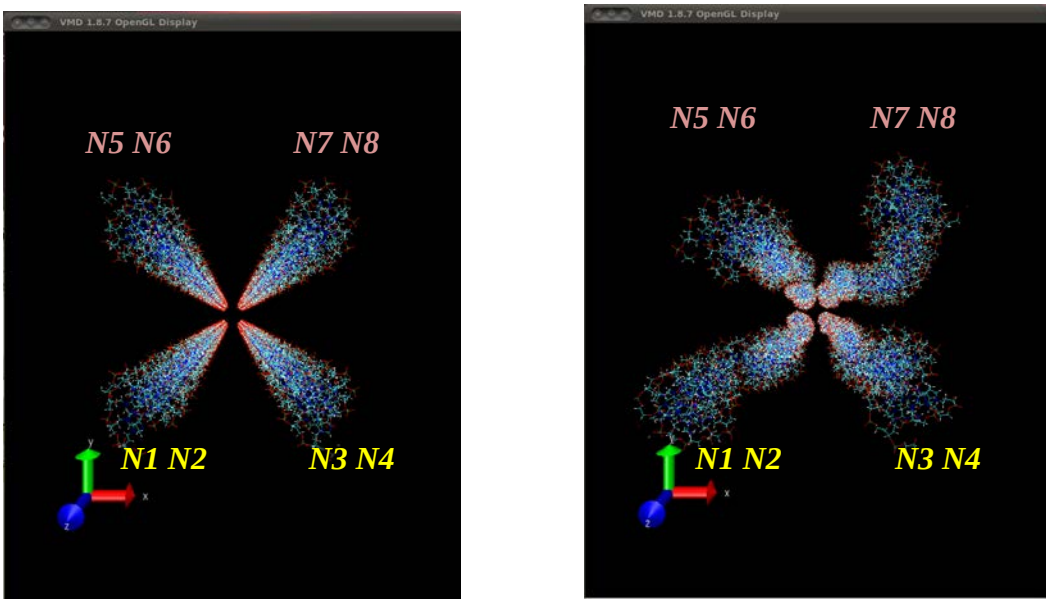


Figure 22: Conformational Change of PhiX-174 associated by Segment Name

Segments N1, N2, N3 and N4 correspond to sequence 176-469 and are denoted by yellow text. Segments N5, N6, N7 and N8 correspond to sequence 406-699 and are denoted by mauve text. Although obvious segregation can't be determined from visual observation a useful TCL command "measure center" provides a more detailed report. The simulation trajectory files were loaded into VMD (Virtual Molecular Dynamics tool) and the center of mass was calculated for each helix on the first frame of the trajectory and the very last frame of the trajectory. Using the 3D coordinates for the centers of mass an Excel spreadsheet was used to then calculate the initial distance and the final distance between each molecule using the displacement calculation below.

$$\text{distance} = \sqrt{dx^2 + dy^2 + dz^2} = \sqrt{(startx - endx)^2 + (starty - endy)^2 + (startz - endz)^2} \quad \textbf{(7)}$$

B22 =SQRT((B7-C7)^2+(B8-C8)^2+(B9-C9)^2)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1							Relative Movement							
2														
3		N1 N2	N3 N4	N5 N6	N7 N8			N1 N2	N3 N4	N5 N6	N7 N8			
4	StartX	0.68009	39.3147	0.75838	39.399		N1 N2							
5	StartY	0.65081	0.56574	39.4851	39.34		N3 N4	2.356082						
6	StartZ	495.032	495.193	495.16	495.182		N5 N6	0.663442	1.119625					
7	EndX	0.28124	41.2601	1.94312	41.0095		N7 N8	1.376927	0.121355	0.441156				
8	EndY	0.32765	0.0608	39.7907	38.9203									
9	EndZ	494.785	493.817	494.792	495.476									
10														
11	Starting Distance Between Segments						Relative Movement Between Similar and Dissimilar Segments							
12														
13		N1 N2	N3 N4	N5 N6	N7 N8		Segment Type Similar			Segment Type Dissimilar				
14	N1 N2	0	38.63508	38.83459	54.73597									
15	N3 N4	38.63508	0	54.78423	38.77434		2.356082			0.663442				
16	N5 N6	38.83459	54.78423	0	38.64087		0.441156			1.376927				
17	N7 N8	54.73597	38.77434	38.64087	0					1.119625				
18										0.121355				
19	Ending Distance Between Segments						AVG			AVG				
20		N1 N2	N3 N4	N5 N6	N7 N8		1.398619			0.820337				
21	N1 N2	0	40.99116	39.49804	56.1129									
22	N3 N4	40.99116	0	55.90385	38.89569		Average movement	1.013098						
23	N5 N6	39.49804	55.90385	0	39.08203									
24	N7 N8	56.1129	38.89569	39.08203	0		NORMALIZED Average Relative Movement Between Similar and Dissimilar Segments							
25														
26	Starting Distance Between Segments (XY ONLY)						Segment Type Similar			Segment Type Dissimilar				
27														
28		N1 N2	N3 N4	N5 N6	N7 N8		1.342984			-0.34966				
29	N1 N2						-0.57194			-2.39002				
30	N3 N4	38.63474								0.106528				
31	N5 N6	38.83438	54.78422							-1.13445				
32	N7 N8	54.73576	38.77434	38.64086										
33							0.385521			-0.9419				

Figure 23: Spreadsheet Calculating COM Movement

The initial distance was subtracted from the final distance to obtain the relative movement in Angstroms of each molecule with respect to each of the other 3. A positive value indicates movement away from each other and a negative value represents closure. The results were tabulated and superimposed on a graphical representation of the virtual system as positive and negative vectors that correspond to separation/closure. The illustration is depicted below in Figure 24.

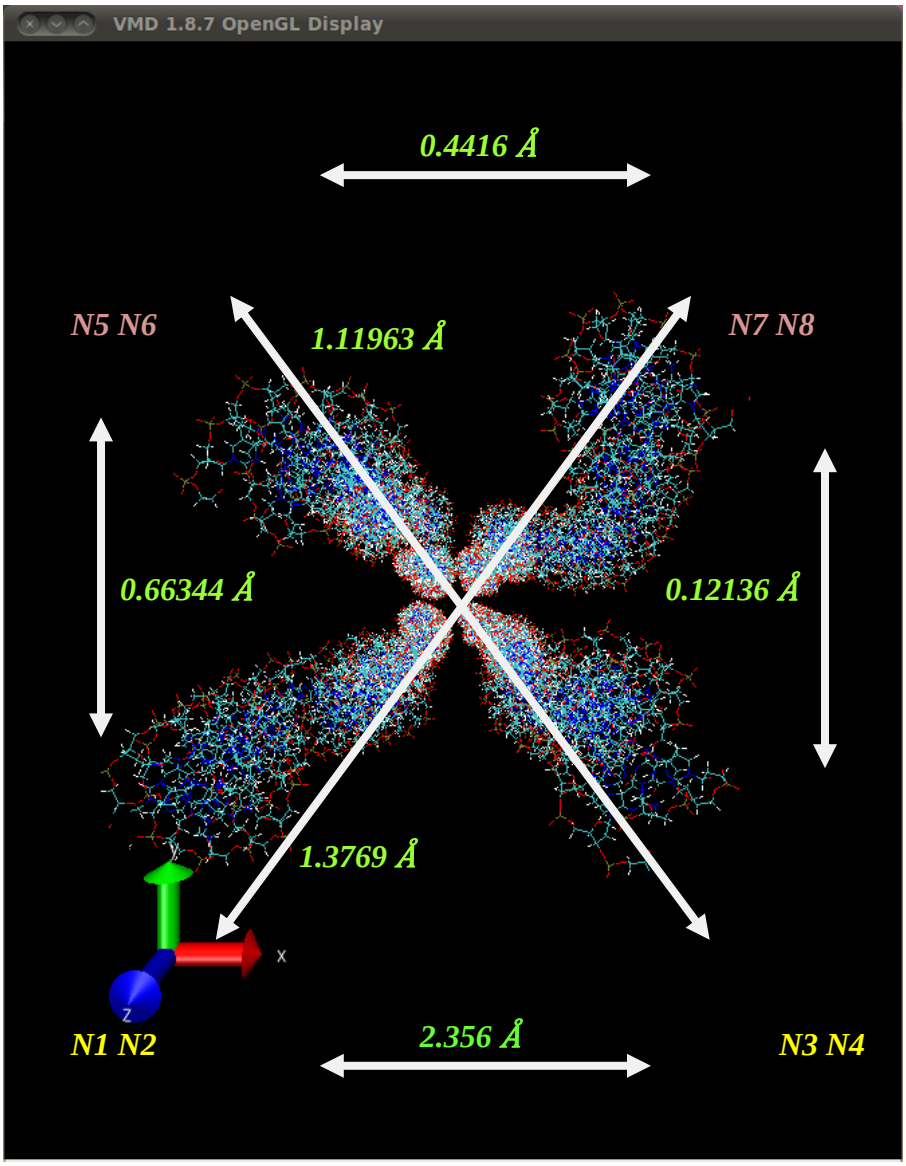


Figure 24: Relative Movement during 2 ns Simulation Time

By visual observation of all positive values one can see that during 2ns of MD simulation time *no-closure was observed between any molecules*. This over-simplified observation requires further statistical analysis of our two tail hypothesis.

Statistical Analysis of closure for Experiment #1

In order to gain more insight to this data as well as formulate a statistical test of our inference the interim behavior between the first and the last positions was examined closely. The positional data taken from the trajectory files was again converted into incremental movements. Instead of just the first and the last positions 47 different positions equally spaced in time from the beginning to the end of the simulation were calculated and tabulated in Excel. The number of data points was the result of two considerations. First the Central Limit Theorem and the Strong Law of Large Numbers (Law & Kelton, 2000) tells us that a sample size of 30 or more will tend to be normally distributed. Second the minimum simulation segment (stop and restart segment) was 100 data outputs meaning the largest even output step is half that or 50. Given these 2 constraints the most convenient number of iterations turned out to be 47. After performing the calculations the resulting data set looked like Table 1. The entire data set is included in the Appendix.

Table 1: Six Categories of Positional Data

	A	B	C	D	E	F	G
1	Step	12DELT32	12DELT56	12DELT78	34DELT56	34DELT78	56DELT78
2	1	0.28757	0.316363	0.29762	0.406964	0.299196	0.093495
3	2	0.166592	0.085073	0.154551	0.076688	0.10685	-0.03151
4	3	0.222058	0.130968	0.249749	0.147499	0.136212	0.07417
5	4	-0.0403	0.112229	-0.02059	0.121237	0.050342	0.018387
6	5	0.061533	-0.00461	0.032386	0.228722	0.165395	0.147079
7	6	-0.02901	0.065387	0.291235	0.126868	0.054696	0.50009
8	7	0.19399	0.064281	0.11994	0.056203	0.039473	-0.04864
9	8	0.055158	0.060403	-0.08437	0.214228	-0.1425	0.212503
10	9	-0.01792	0.251866	0.127901	0.156992	0.239143	-0.07083
11	10	-0.07273	-0.25458	-0.053	-0.10431	0.026913	0.080718
12	11	0.096935	0.251195	0.144863	0.079596	-0.02229	-0.0093
13	12	-0.05819	-0.01044	-0.08592	0.116929	-0.10162	0.218979
14	13	0.057616	0.194144	0.058976	0.181762	0.081312	0.003597
15	14	0.01426	0.204095	0.100002	-0.05545	0.054332	-0.21291
16	15	0.195778	-0.00194	0.087825	0.250952	0.100816	0.186126
17	16	0.188237	0.142695	0.204259	-0.08462	-0.18711	0.026835
18	17	-0.14998	-0.01496	-0.10242	0.164633	0.233521	0.017784
19	18	-0.08322	0.030187	-0.23051	0.111553	-0.15402	0.044728
20	19	0.074143	-0.12649	-0.2039	0.077896	-0.19807	0.07868

The typical column heading “12DELT34” indicates an incremental delta between molecule N1N2 center of mass and molecule N3N4 center of mass. This nomenclature is consistent through the other 5 columns. The column heading “Step” indicates a time step during the simulation. From a statistical standpoint we can look at this as a single independent variable ‘step’ and 6 dependant variables ‘relative movements’. The data can be considered statistically “interval” data because it represents a linear measure of Angstroms that remains consistent throughout the scale. If the means of the dependant variables can be tested with a null hypothesis that they are zero we can infer that the results of that test would also apply to the

accumulated positional change over time. This is because the accumulated displacement over time is a direct function of the tested means.

Referring to guidelines published by the UCLA Statistical Consulting Group (UCLA: Academic Technology Services, 2011) We can accomplish this with a straightforward t-test as long as the data is independent and identically distributed (IID). Since there are 47 data points we can safely assume it is identically distributed normal because the Central Limit Theorem tells us that sample sizes of 30 or more tend to normal distributions. (Ludford) The question of independence must be considered carefully. The intuitive meaning of independence tells us that the value of one data point in one data set will have no effect on the probable value of a data point in a different data set. Thinking of the Phi X 174 molecular system, the movement of any of the molecules *relative to an adjacent molecule* can be considered independent. For example, if N1N2 moves relative to N5N6 the movement is only very slightly related to N1N2 moving relative to N3N4. They can move independently without significantly influencing each other. On the other hand, movement relative to diagonal molecules exhibits a much greater effect on adjacent movements. The relationship would be equivalent to the arc length of a 90 degree sweep for a given radius versus the arc length of a sweep of less than 1 degree for the same radius. For example, if N1N2 moves relative to N5N6 the diagonal distance between N1N2 and N7N8 is significantly affected. For this reason only the four adjacent movements should be tested assuming IID data.

In addition to testing the means of adjacent movements to be sure they are not zero it would be very informative to know if they are equal to each other. Three different sources were examined for guidance in determining an appropriate test for this point of interest, two guideline

papers (McCrum-Gardner, 2008) (UCLA: Academic Technology Services, 2011) and a web critique of a previous application where the data structure is similar to this data (Ludford). After consideration of these guidelines it was determined that our movement data can best be described as independent normally distributed repeatedly measured interval data in four categories. The test chosen is the One Way Repeated Measures ANOVA test. The data consists of matched sets of sample members where each set has the same number of sample members and the members of each set represent uniquely different conditions. When sample members are matched in this way, measurements across conditions can be appropriately treated just like repeated measures in a standard repeated measure ANOVA.

The widely available statistical analysis program SAS 9.1.3 with service pack 4 was used to perform both tests. Beginning with the One Sample t-test the spreadsheet shown above in Table 1 was imported directly into SAS. The test was applied to each of the adjacent categories using the following syntax:

```
proc ttest data = "Phix174.dispdata" h0 = 0;
  var _2DELT32;
run;
```

The h0=0 tells SAS to test the null hypothesis that the mean N1N2DELTN3N4 is not significantly different from zero.

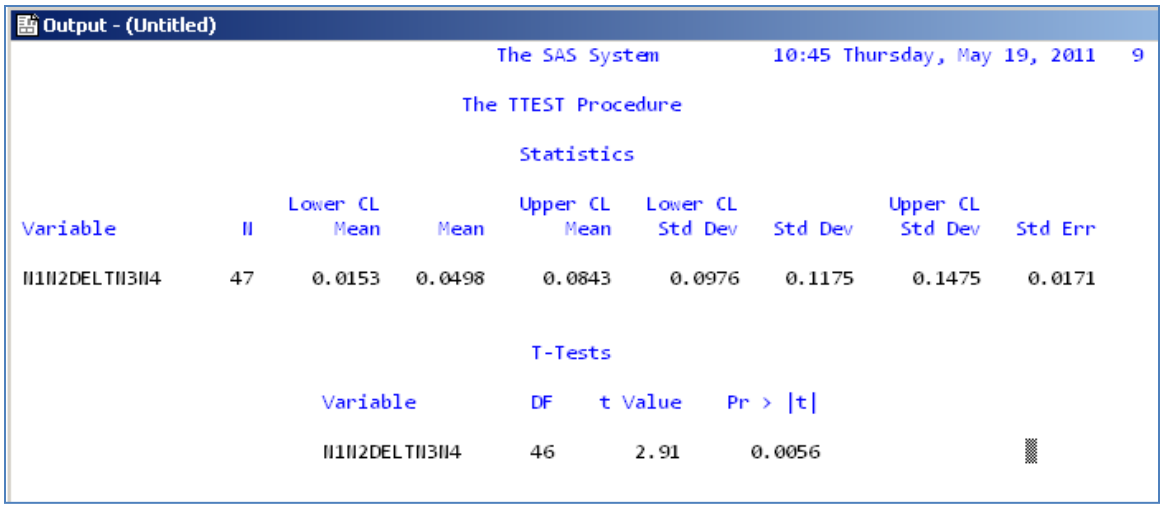
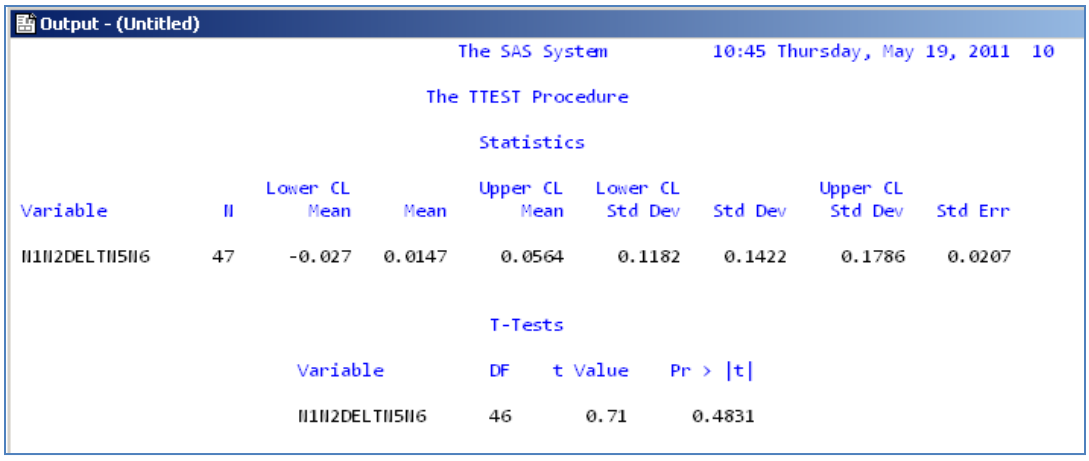


Figure 25: Example SAS Output Screen for N1N2 rel N3N4

The resulting SAS output in Figure 25 gives us a t value of 2.91 and a P value of 0.0056. This means that if the null hypothesis were correct and the population mean was not significantly different from zero there is a less than 6 in 1000 chance that 't' would be bigger than 2.91 using data from the same population. Stated in another way, with a chosen significance level of 0.05 (95%), we observe that p=0.0056 being less than 0.05 provides significant reason to accept the alternative hypothesis that the mean is NOT zero. We can extrapolate this conclusion to an accumulation of movements from this same population inferring that it too is significantly different from zero. More importantly, with regard to the general research hypothesis, it is non-zero and *positive.* In practical terms this means that molecular segments N1N2 and N3N4 drifted *away* from each other contradicting the expected outcome for similar sequence molecules. The results of the t-test for the remaining 3 sides of the "box" are equally unexpected. The following 3 figures depict the remaining adjacent movement tests.



Output - (Untitled)

The SAS System 10:45 Thursday, May 19, 2011 10

The TTEST Procedure

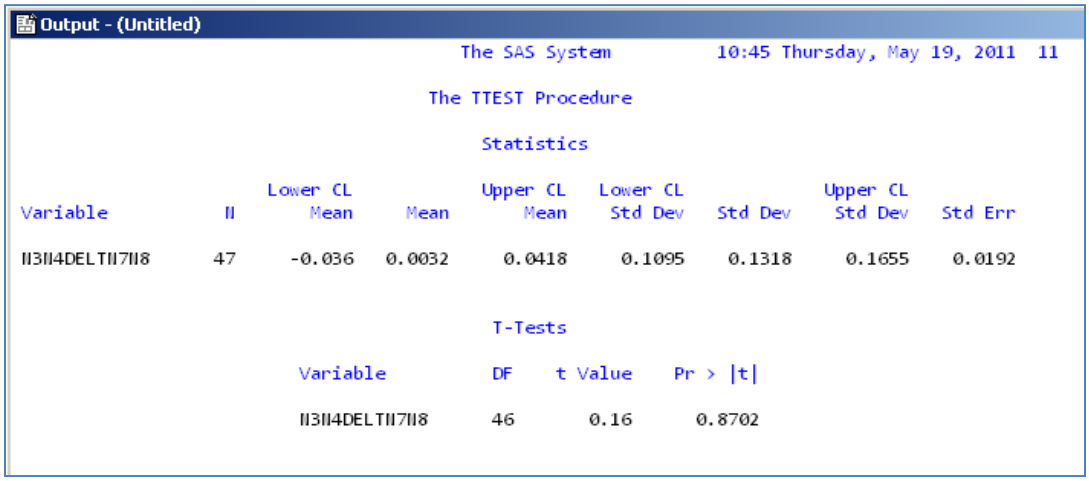
Statistics

Variable	N	Lower CL Mean	Mean	Upper CL Mean	Lower CL Std Dev	Std Dev	Upper CL Std Dev	Std Err
N1N2DELTN5N6	47	-0.027	0.0147	0.0564	0.1182	0.1422	0.1786	0.0207

T-Tests

Variable	DF	t Value	Pr > |t|
N1N2DELTN5N6	46	0.71	0.4831

Figure 26: SAS Output Screen for N1N2 relative to N5N6



Output - (Untitled)

The SAS System 10:45 Thursday, May 19, 2011 11

The TTEST Procedure

Statistics

Variable	N	Lower CL Mean	Mean	Upper CL Mean	Lower CL Std Dev	Std Dev	Upper CL Std Dev	Std Err
N3N4DELTN7N8	47	-0.036	0.0032	0.0418	0.1095	0.1318	0.1655	0.0192

T-Tests

Variable	DF	t Value	Pr > |t|
N3N4DELTN7N8	46	0.16	0.8702

Figure 27: SAS Output Screen for N3N4 relative to N7N8

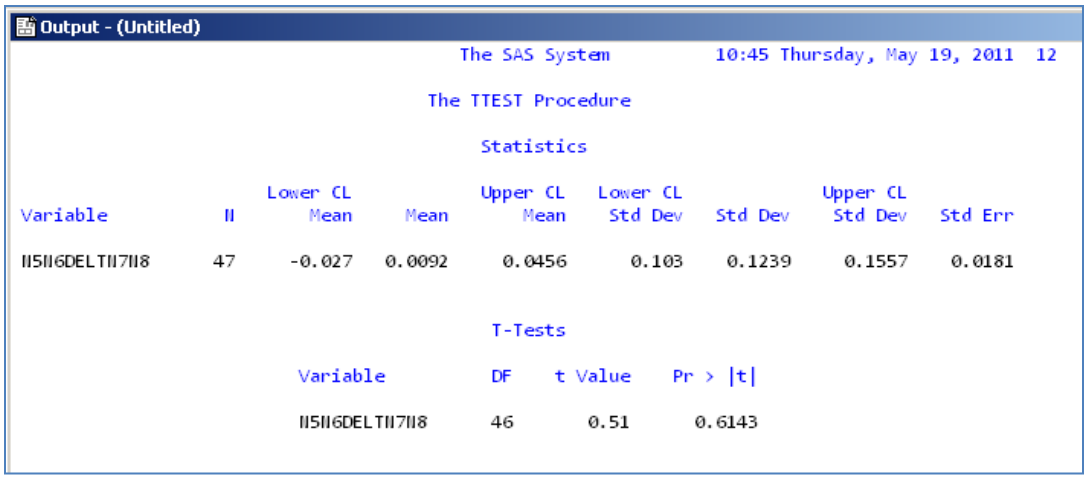


Figure 28: SAS Output Screen for N5N6 relative to N7N8

Contrary to expectations, with P values of 0.4831, 0.8702 and 0.6143 ALL greater than 0.05, the remaining 3 adjacent movement means are NOT significantly different from zero.

The One Way Repeated Measures ANOVA test is run with this command syntax:

```
proc glm data = PhiX174.DispData;
  model N1N2DELTN3N4 N1N2DELTN56 N3N4DELTN7N8 N5N6DELTN7N8 = ;
  repeated Step ;
run;
quit;
```

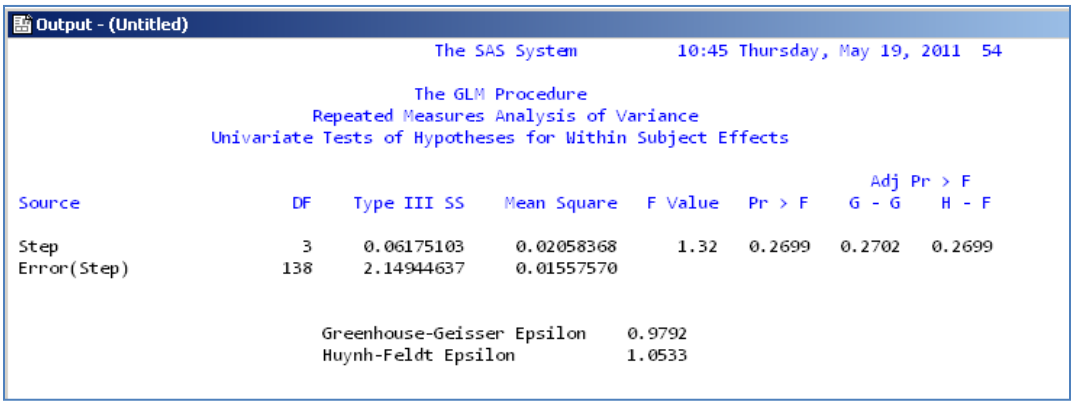


Figure 29: SAS Output Screen for Repeated Measures ANOVA

For an alpha decision point of 95% the calculated P value would have to be less than 0.05 to confidently reject the null hypothesis. From the calculated P value of 0.2699 we cannot reject the null hypothesis that the 4 means are equal.

In summary, with 3 of the 4 means testing insignificantly different from zero and the 4 means together testing insignificantly different from each other one can only conclude the expected results of directed motion are not reproduced by this simulation data.

Although this simulation presented no evidence to support $\mathbf{H}_{(\text{Simulate Observed Closure NULL})}$ that current generation molecular models can simulate DNA segregation like that observed by the Imperial College team, or $\mathbf{H}_{(\text{Resonance Causes Closure NULL})}$ that resonance causes closure between DNA molecules, it should be pointed out that the high cost in time and money for data of this type resulted in a small-scale research plan that may simply have been in-sufficient to adequately investigate these hypotheses. Re-running the simulations with longer simulation run-times, multiple geometric configurations and larger simulation spaces may produce results in support of the hypotheses. In summary the results of Experiment #1 indicate we must reject $\mathbf{H}_{(\text{Simulate Observed Closure NULL})}$. Because we were unable to simulate closure at all we are unable to satisfactorily test the follow on hypothesis $\mathbf{H}_{(\text{Resonance Causes Closure NULL})}$ with the resources that were available to us. Steps that can be taken to more thoroughly test $\mathbf{H}_{(\text{Simulate Observed Closure NULL})}$ subsequently allowing testing of $\mathbf{H}_{(\text{Resonance Causes Closure NULL})}$ are outlined in the Conclusions section.

Experiment #2 Resonance Results

To accomplish Experiment #2 a total of five molecular systems were constructed and run through molecular dynamic simulation for more than 1560 cpu hours completing the proposed investigation. All together more than 64 NAMD format simulation scripts and 64 NAMD format simulation batch files were written to perform nearly 50 MD simulations not including practice runs and verification runs. Because pressures are of primary interest to this research the simulation task using NAMD is nearly doubled. The internal complexities of on-the-fly pressure calculation and the resulting burden on simulation speed causes NAMD to output only part of the desired total system pressures even when profiling is activated. To obtain the complete pressure variable for an MD run the simulator must be run twice, the second run calculating and outputting only the Ewald component of system pressures. The two outputs must then be summed offline by separate programming methods to get the desired result. (Bhandarkar et al., 2008) After the actual simulations were complete (including pressure calculation runs) 56 custom programs were written to perform this raw data parsing and manipulation generating 252 summary data files. The PERL (Practical Extraction and Reporting Language) computing environment was used for general data parsing and manipulation. PERL script templates were written titled “Parse_Ewald_Pressures.pl”, “Parse_Runtime_Pressures.pl” and “Sum_Runtime_Ewald_Pressures.pl” totaling 1286 lines of PERL code with 775 of those lines in just the summing script. Each PERL template was then copied to all 5 molecular system directories and customized to operate on a specific data set. Finally 12 additional Perl scripts were written to perform search and match functions on the summary data files from which

summary spreadsheets were created by hand. At the time of this writing the sum total of simulation data now exceeds 152GB and continues to grow.

To accomplish the final statistical analysis of the raw data the MATLAB® (Moler, 2004) numerical computing environment was used. More than 70 MATLAB® scripts were written during the data analysis phase for data parsing and statistical characterization. All MATLAB® development culminated in 5 final program templates totaling over 800 lines of code with 537 lines in the analysis program alone. These 5 final analysis template programs were copied to each of the 5 molecular system directories and customized to operate on the specific data for each system generating a unique solution set for each system. Each solution set consists of several MATLAB® graphic figures, one raw data output text file and 1 statistical summary text file.

Experiment #2 necessarily took place in two steps. The first step was to construct and run 4 initial test systems using the 4 unique configurations proposed. The 4 initially proposed systems were constructed and MD simulations were run in accordance with the same best practices used for the Phi X 174 system. Graphical representations of each of the 4 systems are illustrated in Figure 30 through Figure 33. The red and blue spheres represent nucleic acid chains and the larger yellow\light blue spheres represent Na^+ and Cl^- ions (salt) in the solvent. Below are illustrations of the actual test systems that are reasonably scaled with the exception of the size of the nucleic acids and ions being exaggerated to make them more visible among the water molecules.

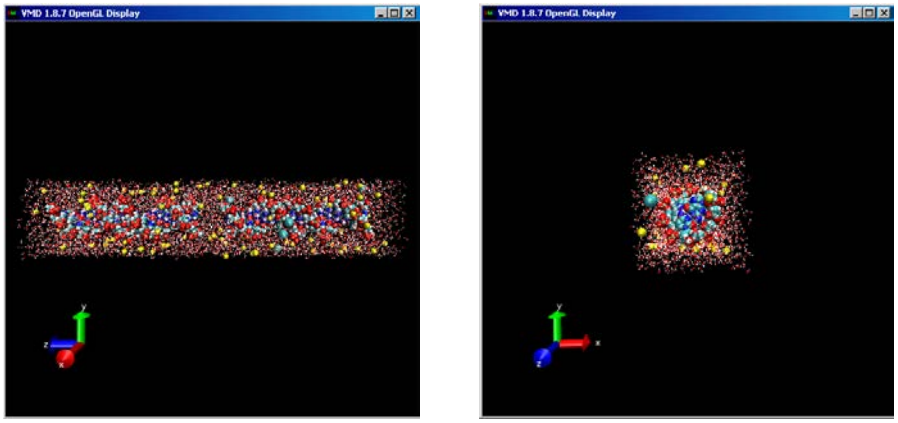


Figure 30: End-to-end Linear Configuration

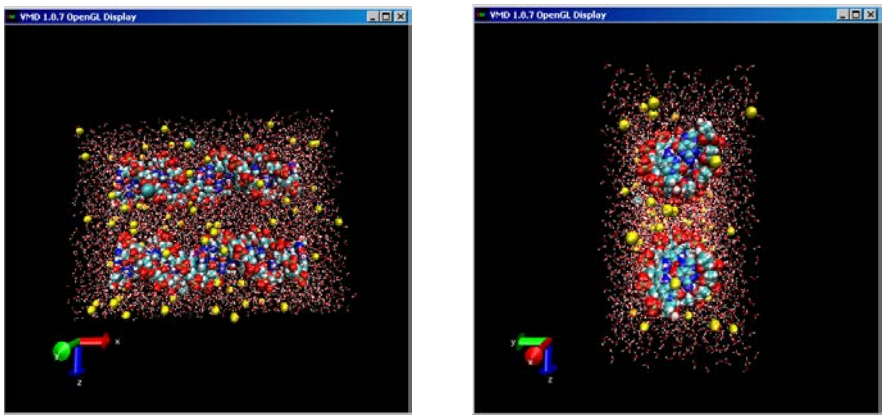


Figure 31: Parallel Configuration

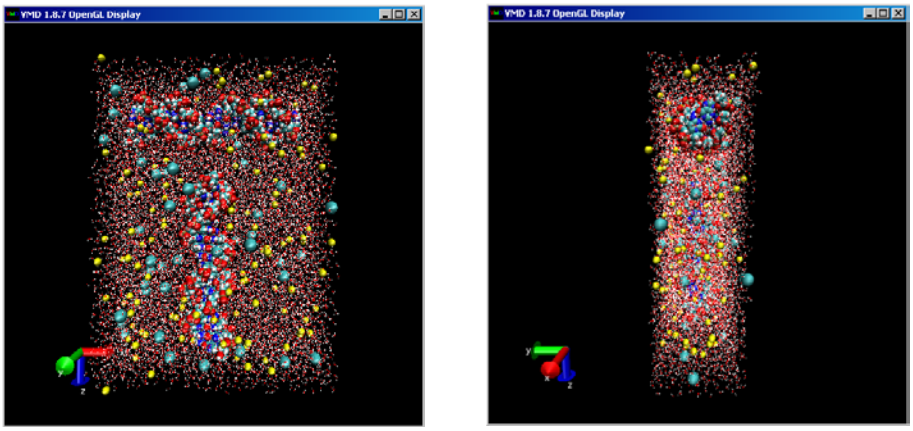


Figure 32: Perpendicular "T" Configuration

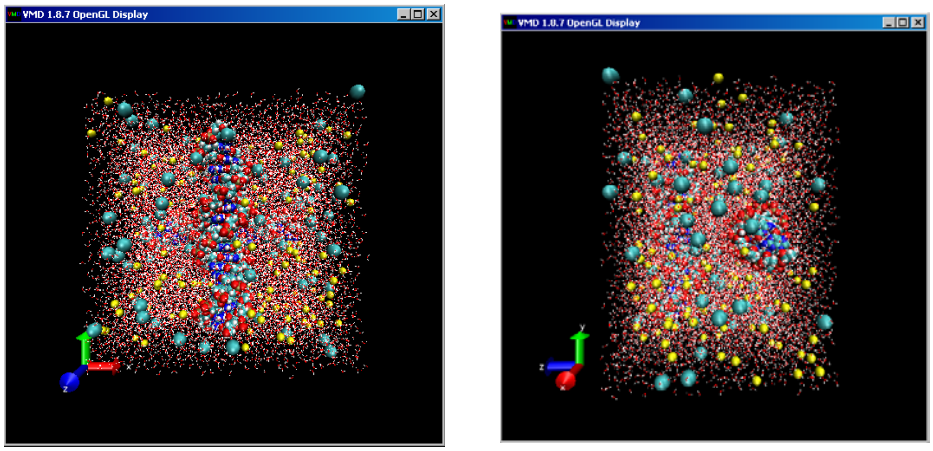


Figure 33: Skew Configuration

The total run time for all systems ended up being determined by movement of the DNA within the water boxes. At the beginning of each simulation the DNA molecules were centered exactly within the solvent box. As MD simulations progressed the DNA would move randomly and eventually at some point exceed the confines of the water box. When this happens the system neither becomes un-stable nor is the simulation terminated because the NAMD simulator "wraps" molecules into periodic replicas maintaining simulation integrity. The details of this technique as well as the effect on various simulation variables are beyond the scope of this research. The interested reader can find many publications addressing this topic beginning with the simulator's original publications. (Phillips et al., 2005) With respect to potential structure related resonation this becomes an unacceptable situation due to loss of fidelity (i.e. each system can no longer be considered an accurate model after the DNA had extended beyond the limits of the box). Figure 34 is an illustration of this condition depicted for the end-to-end linearly configured system below.

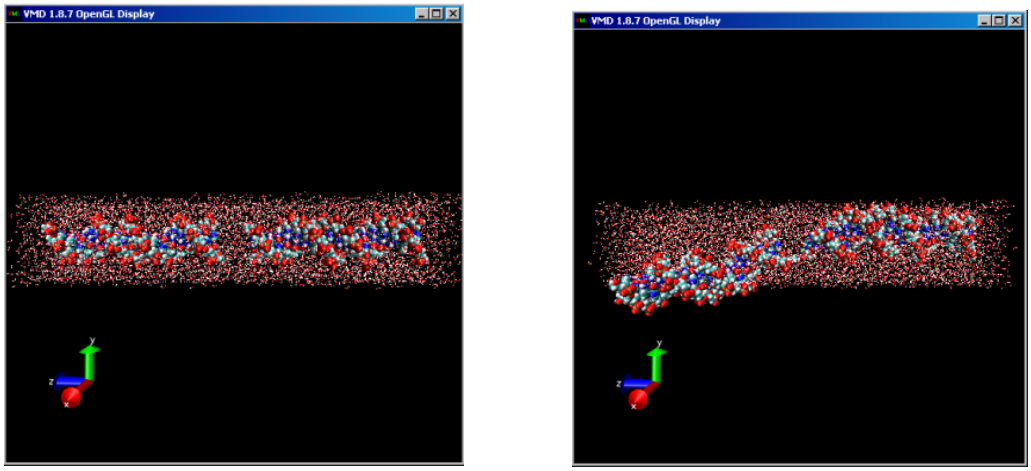


Figure 34: End-to-End Linear Configuration at start of MD and at 2us

In the interest of consistency and to avoid complications from molecular break out all simulations were limited to a total of 2us of simulation run-time. The system configured in an end-to-end linear fashion was originally intended to be re-configured with a contrasting sequence but if one of the other systems would have been more appropriate a determination needed to be made at this point. With the first 4 molecular system simulations complete an assessment of the data was needed before construction of the 5th and final system.

Statistical Analysis for Experiment #2

To investigate $\mathbf{H}_{(\text{Resonance NULL})}$, $\mathbf{H}_{(\text{Harmonized Resonance NULL})}$ and $\mathbf{H}_{(\text{Interacting Harmonized Resonance NULL})}$ we will apply a Fourier Transform to simulated pressures generated by MM algorithms. The result of the transform will be a frequency "spectrum" of the real-time pressure data. This was accomplished using the MATLAB® FFT (Fast Fourier Transform) which is simply a high-speed algorithm implemented in the MATLAB® computational environment that returns a Fourier transform in a timely fashion (i.e. Fast). With the results of the FFT in hand it becomes necessary to make a statistical assessment of what the spectrum is telling us.

At the time of this writing the application of Fourier analysis to intermolecular pressures simulated by MM algorithms is unprecedented. Because of this one must proceed carefully and establish a few very basic principles to build upon. From a very general perspective our interest in the FFT data can be summarized by one question, is there any statistically significant frequency content at all in the data?

As with Experiment #1, descriptive statistics alone will not inform us of this. A parametric test of significance of some sort is needed. Initially two common analysis techniques, data decimation (re-sampling) and windowing (Hanning windows), were considered. These techniques turned out to be un-helpful for two main reasons. First they both require *prior knowledge* of periodic content to be implemented correctly. For example, if an inappropriate Hanning window function is applied to raw FFT data, errors can be introduced to the amplitude, frequency and overall shape of the FFT transform depending upon if the data contains only periodic data or a combination of periodic and non-periodic data (DC or Direct Current from Signal Analysis jargon). (Williams, 2004) Data re-sampling will also corrupt FFT data through

aliasing if the re-sample interval is implemented without appropriate low pass filtering. Like Hanning window functions, choosing the most appropriate re-sample interval again benefits from prior knowledge of frequency content within the data. (Mercer, 2001) If applied correctly, these techniques would be informative regarding specific frequency and magnitude information; however, they do not speak directly to the general question of the basic existence of periodic behavior beyond normal modes of water and so were not considered further. As mentioned, descriptive statistics alone is insufficient for our needs but clear descriptive methods are still required for any kind of testing, so that is where we have to start.

Again, published research in the literature addressing the use of Fourier Analysis directly on molecular dynamic simulation data is scarce so our analysis therefore has to begin with first principles. Statistically describing Fourier Coefficients must begin with several assumptions. First we remember that molecular motion in a liquid is a random process known as Brownian motion that is Gaussian Markov in nature (Pitman, 2003) This research is based on the assumption that there is non-random periodic behavior occurring in addition to the usual normal random motion that will manifest itself as periodicity in the inter-molecular pressure data. In engineering jargon these variables are commonly referred to as signal and noise. This leads us to one of the most important assumptions we are going to make about the system. That assumption is that the signal we are looking for and the noise in the system are linearly *additive*. This means that the raw data represents the linear sum of the periodic signal and the random noise. The second important assumption we are going to make is that the noise portion of the data represents a sample that is independently drawn from a process that has a zero mean and a variance σ^2.

Statistically this implies that a noise value $noise_j$ will be a random variable that is independent and identically distributed. With respect to our data this means that a sample pressure $pressure_j$ in the data series will consist of a sum of the signal of interest $signal(x_j)$ and random noise $noise_j$. Therefore $pressure_j$ can also be treated as a random variable. Furthermore, the noise will be additive and its mean will be zero. Since the signal is additive and noiseless we can assume that the variance of $pressure_j$ will be equal to the variance σ^2 of the noise. This can be summarized as:

$$pressure_j = signal(x_j) + noise_j \quad \textbf{(8)}$$

$$\overline{pressure_j} = signal(x_j) \quad \textbf{(9)}$$

$$Var(pressure_j) = \sigma^2 \quad \textbf{(10)}$$

Since our pressure data is a sum of our signal (or signals) of interest and a noise component, the Fourier coefficients produced by the MATLAB® Fast Fourier Transform function can be considered estimates of the true Fourier coefficients of our signal of interest. With this in mind the best way to characterize our estimates is in terms of a probability

distribution. If the signal of interest is deterministic then the probability distribution of the Fourier coefficients generated by MATLAB® for D samples of data will depend on the probability distribution of the noise. Again, assuming the noise is Gaussian having a normal probability density of mean μ and variance σ^2 the probability P that the noise component at any moment will lie somewhere between a and b will be given by the area under the normal probability density function operating between the limits a and b. This is stated mathematically like this:

$$\mathrm{P} = \int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-u)^2/2\sigma^2} dx \qquad \textbf{(11)}$$

(Thibos, 2003))

This justifies several more helpful assumptions we can now make relying on probability theory. From the central limit theorem we know that the sum of many independent variables will tend to a Gaussian distribution no matter what the distributions of each variable. Also from probability theory we know that a Gaussian distribution is closed under addition. This tells us that the distribution of a weighted sum of Gaussian variables will also be Gaussian. Since our data consists of the sum of a deterministic noiseless signal component and a Gaussian noise component the Fourier Coefficients returned by MATLAB® will also be Gaussian! This is very helpful as will be shown.

We need to address one more concept relevant to the analysis, signal *power* or energy per unit time. We know that Fourier coefficients are complex numbers by definition. From

Parseval's Theorem we know that total power in a data vector of Fourier coefficients is the sum of the squared amplitudes of the Fourier coefficients divided by 2 plus the square of the average. (Thibos, 2003) The average (DC term or c_0) is of no interest to us because it informs us of nothing with regard to periodic variation. Therefore, the signal power we are interested in is simply the coefficient amplitude squared over 2. Stated mathematically, the estimated power p_k in the k-th harmonic is:

$$p_k = (\hat{a}_k^2 + \hat{b}_k^2)/2 \qquad \textbf{(12)}$$

From probability theory again we know that a standardized Gaussian variable with zero mean and a variance of 1 (unit variance) squared will be distributed as a chi-squared variable with 1 degree of freedom. In a similar fashion squared standardized Fourier coefficients will also be distributed as chi-squared. Applying this to the Fourier coefficients from our pressure data implies that:

$$\frac{(\hat{a}_k - a_k) + (\hat{b}_k - b_k)}{2\sigma^2 / D} \xrightarrow{\textit{approximately}} \chi_2^2 \qquad \textbf{(13)}$$

(Thibos, 2003)

At last we have a way to statistically test for the presence of periodic variation at particular frequencies. A simple *null* hypothesis can now be formulated that the Fourier coefficient of the

kth harmonic frequency is *zero*. The resulting test statistic was first developed in 1949 by H.O. Hartley and is commonly known as the H statistic. (Hartley, 1949) Under this null hypothesis Equation 13 becomes:

$$\frac{(\hat{a}_k^2 + \hat{b}_k^2)}{2\sigma^2 / D} \xrightarrow[approximately]{} \chi_2^2 \quad \text{(Thibos, 2003)} \qquad \textbf{(14)}$$

Substituting in the signal power previously defined we get:

$$\frac{p_k}{\sigma^2 / D} = \frac{\text{Power in the kth harmonic}}{\text{Average noise power}} \xrightarrow[approximately]{} \chi_2^2 \quad \text{(Thibos, 2003)} \qquad \textbf{(15)}$$

Now that we know that harmonic power will follow a chi-squared distribution when only Gaussian noise is present we can use an F-test to examine the goodness of fit of a Fourier model. The derivation begins by assigning the left side of Equation 15 as the numerator of an F statistic. Since there would be D-3 residual harmonics the total relative power in the residual harmonics would be the sum of R=(D-3)/2 random variables with 2R=D-3 degrees of freedom resulting in:

$$\sum_{j=1}^{R} \frac{p_j}{\sigma^2 / D} = \xrightarrow[approximates]{} \chi_{2R}^2 \quad \text{(Thibos, 2003)} \qquad \textbf{(16)}$$

To obtain Hartley's test statistic we divide each variable by the corresponding number of degrees of freedom forming the ratio:

$$\frac{\frac{p_k}{2\sigma^2/D}}{\frac{1}{2R}\sum_{j=1}^{R}\frac{p_j}{\sigma^2/D}} = \frac{\text{relative power in kth harmonic}}{\text{average relative power in residuals}} \xrightarrow[approximates]{} F_{2,2R} \qquad \textbf{(17)}$$

(Thibos, 2003)

The statistic simplifies to:

$$H = \frac{p_k}{\frac{1}{R}\sum_{j\neq k} P_j} = \xrightarrow[approximates]{} F_{2,2R} \qquad \textbf{(18)}$$

(Thibos, 2003)

Simply put, the null hypothesis that the power in the k^{th} harmonic is zero can be rejected if

$H \succ F_{2,2R}$ (Thibos, 2003) **(19)**

The practical application of this test for a particular significance level (like .1%) would be to obtain the value of the F distribution (F usually from a table) for the desired significance. Then compare this value to the H statistic calculated per Equation 18. If the H statistic is greater than the F-value reject the null hypothesis that the power is zero in that particular harmonic. The significance level 0.1% represents the probability of falsely rejecting the null. In general this analysis will help to determine which harmonics, if any, should be included in a Fourier series

model of a signal. With regard to this research no specific information exists about frequency, the objective being to simply determine if there is any frequency content at all. Therefore, the coefficients will be ranked by their magnitudes in order of greatest to smallest; each coefficient in turn will be tested for significance until any/all significant harmonics have been determined. This provides us with a convenient way to generally characterize an FFT spectrum and statistically test individual coefficients for significance, using power.

Confidence Intervals

The next logical question to address about a test statistic is how confident are we in our characterization? Recalling from basic statistics, a common way to specify confidence bounds about a population mean is to define with some chosen probability alpha (i.e. for alpha of 0.05 there will be less than 5% chance of being wrong) a range

$$\bar{x} - A \leq u \leq \bar{x} + A \qquad \textbf{(20)}$$

within the true population mean will fall within. All we need is the value of A and we will have our interval. To find this we start by recalling the definition of Student's t-statistic:

$$t = \frac{|\bar{x} - u|}{s/\sqrt{N}} \qquad \textbf{(21)}$$

Where t is a standardized sample mean, s is the sample standard deviation and $s/\sqrt{N}$ is the standard error of the mean. It will have the t-distribution with N-1 degrees of freedom. In general the t-distribution looks like this:

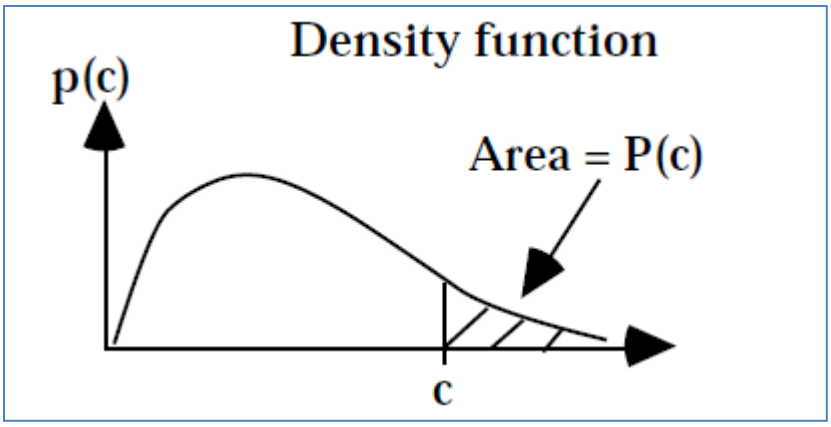


Figure 35: Probability Density Function

In Figure 35 above p(c) is the probability density function. Below is a graph of 1 minus the cumulative probability distribution or P(c). P(c) is also the area under the probability density function past a given c. The precise value of c where P(c) is equal to or less than our chosen alpha (in this case 5%) is dependent on D but, for very large samples we know that c is approximately 2. This can be interpreted generally as "the probability of t being greater than 2 is around 5%".

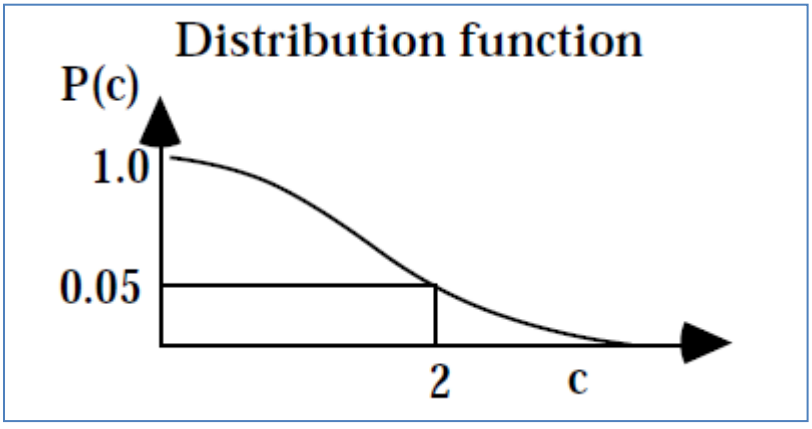


Figure 36: Probability Distribution Function

Referring back to Equation 21 we can infer:

$$\text{Prob}\left(\frac{|\bar{x}-u|}{s/\sqrt{N}}\right)=5\% \qquad \textbf{(22)}$$

Restating the inequality portion of the equation we obtain:

$$\text{Prob}\left(\bar{x}-2s(\bar{x}) \prec u \prec \bar{x}+2s(\bar{x})\right)=95\% \qquad \textbf{(23)}$$

Recalling that Equation 17 informed us that Hartley's ratio of harmonic power to the power in the residuals exhibits the F-distribution under a null hypothesis restriction, we can eliminate the null hypothesis restriction and substitute a broader form of the numerator and get the following:

$$H=\frac{(\hat{a}_k-a_k)^2+(\hat{b}_k-b_k)^2}{\frac{1}{R}\sum_{j=1}^{R}P_j}=\xrightarrow[approximates]{}F_{2,2R} \qquad \textbf{(24)}$$

In an analogous fashion to Equation 22 we therefore have:

$$\text{Prob}\left(\frac{(\hat{a}_k - a_k)^2 + (\hat{b}_k - b_k)^2}{\frac{1}{R}\sum_{j=1}^{R} P_j} \geq F_{2,2R} \right) = 5\% \qquad \textbf{(25)}$$

The best way to illustrate the application of Equation 25 to a Fourier Coefficient is to first remember that they are complex numbers with real and imaginary components. To apply this boundary we simply draw a circle centered at $(\hat{a}_k, \hat{b}_k)$ with a radius ρ giving us $\frac{F_{2,2R}}{R}\sum_{j=1}^{R}\rho_j$ that we can inspect directly. For a given alpha of 0.05 we can now state with 95% confidence that the true values of (a_k, b_k) are contained within that circle. Most importantly, if the circle contains the origin, we know that the power in that particular harmonic is not significantly different from zero.

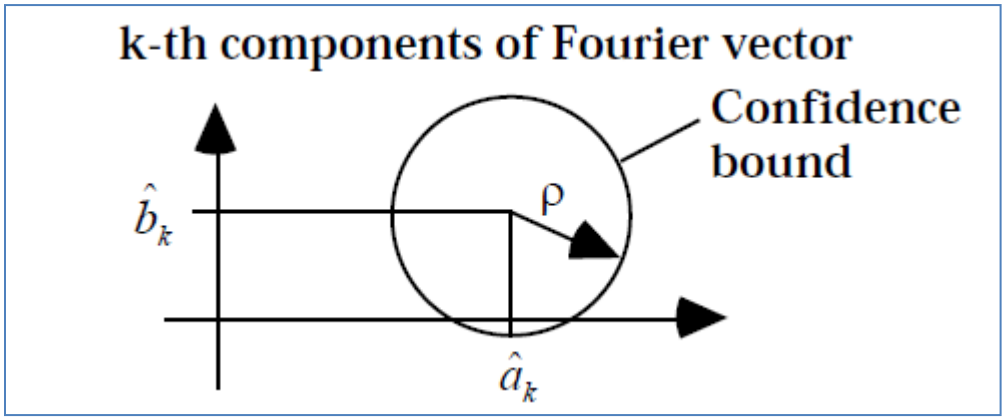


Figure 37: Circular Confidence Boundary for Fourier Coefficients

Now that we have a method of hypothesis testing with confidence intervals for Fourier Coefficients we must consider one more important characteristic of the output data before we can address the question of how to make an assessment of that data. The simulator output is conveniently recorded as pressure tensors with individual values for the X, Y and Z components of system pressures. This was taken into account during molecular system design resulting in all systems being laid out so that the sought after interaction would occur along the Z axis. Although a composite variable of absolute pressure is intriguing, it can reasonably be assumed that no significant pressure contributions will come from the X or the Y dimensions in any of the four systems because the 4 surfaces bounding the solvent slabs between DNA molecules do not exhibit perpendicular exposure to the DNA. In other words, the DNA molecules can only be 'seen' from the plane orthogonal to the Z axis. From the midpoint slab, the X and Y axes will never intersect the DNA therefore any pressure components resulting from DNA-DNA interaction will not be traveling along either the X or the Y axes. With this concept in mind we can focus all our attention on the Z dimension pressure components.

Assessment of Initial Experiment #2 Results

The Hartley statistic described in the *Statistical Analysis* section was implemented in the previously mentioned MATLAB script "Hartley_Combo_Final.m" , located in the Appendix. In addition to the statistical significance test that has been applied to every coefficient returned by the transform an intuitive non-parametric test was applied to 2 summary variables from each run. By taking the total number of coefficients for each system and subtracting the number that tested significant we obtain 2 conditions for each result. By considering each system a group and the

expected outcome a group we obtain 2 counts (or groups) for each condition. We can then construct a standard 2 x 2 contingency table from these 4 data points and test for goodness of fit.

Chi-Square test with Yates correction

As discussed in earlier sections we remember that the basic Chi-Square test is a good method of finding the approximate probability of experimental outcomes arising by chance alone. The basic equation is:

$$\chi^2 = \sum \frac{(\text{Observed Frequency - Expected Frequency})^2}{\text{Expected Frequency}} \qquad \textbf{(26)}$$

Using a 2 x 2 contingency table we can use this equation to test the differences between 2 actual samples. The difficulty with a 2 x 2 table is that it is a very small data set. To account for this the 'Yates correction for continuity' can be used. This takes into account the uncertainty introduced by small samples that can cause an erroneous conclusion that a difference exists when it does not. Employing the Yates correction makes for a more conservative test decreasing the likelihood of a Type I error. This is easily accomplished by subtracting 0.5 from the absolute value of the numerator. The equation then becomes:

$$\chi^2 = \sum \frac{(\left|\text{Observed Frequency - Expected Frequency}\right| - 0.5)^2}{\text{Expected Frequency}} \qquad \textbf{(27)}$$

The practical matter of accomplishing these calculations turned out to be challenging. For a proper analysis it is desirable to see the basic Chi-square value, the Yates correction for

continuity value, and the p value for each of these runs. Getting all these variables easily from the single statistical package that I had access too was not feasible. The basic Chi-square test function within SAS 9.1 does not return values for the p value or Yates Correction, only basic Chi Square. The calculations were performed again with the popular MINITAB 16.1.0 and Chi-square with a p value was returned but again no value for Yates correction is provided. The MATLAB crosstab function returned the same. In the end the easiest and most concise results were obtained from 2 internet based statistical calculators. (Preacher, 2011) (GraphPadSoftware, 2011) The two calculators corroborated each other and were cross checked manually with Excel. The following tabular results were then generated for the 4 original systems.

Table 2: End-to-End, Linearly configured, Identically Sequential, molecule pair Frequency/Power Results Z axis

----------					
Analysis for data file: LinearColumn01z.txt					

Total Number of Coefficients Positive Side of Spectrum					
49997					
Total Number of Significant Coefficients Positive Side of Spectrum					
975					
Percentage of Total that are Significant					
1.95214					
Power in Significant Coefficients div Total Power*100					
22.77335					
Upper and Lower Frequency Limits Returned by Transform					
Lower Frequency	Upper Frequency	Sampling Frequency			
5.000E+008	2.500E+013	5.00000E+013			
F2,2R	Alpha				
6.91	0.001				
Top 100 Coefficients Ranked by Magnitude of Power (DC listed First)					
Hz	Hartley	Power	Hz	Hartley	Power
0.000E+000	n/a	90.5712			
1.000E+010	202.138	201.1453	3.400E+010	22.723	22.6923
1.500E+010	56.790	56.6752	1.265E+011	22.517	22.4865
1.250E+010	54.479	54.3712	3.720E+011	22.388	22.3581
6.250E+010	54.032	53.9260	4.500E+009	21.895	21.8665
2.100E+010	50.488	50.3925	9.666E+011	21.455	21.4265
2.500E+009	44.303	44.2248	1.225E+011	21.421	21.3932

5.000E+009	42.501	42.4269	9.906E+011	21.305	21.2771
4.335E+011	39.655	39.5880	3.465E+011	21.290	21.2617
7.501E+009	39.090	39.0252	1.365E+011	21.240	21.2126
1.750E+010	38.024	37.9615	4.950E+010	21.212	21.1845
6.550E+010	36.668	36.6084	1.325E+011	21.212	21.1843
3.000E+009	35.226	35.1697	3.600E+010	21.128	21.1002
1.950E+011	34.738	34.6833	1.859E+012	20.996	20.9684
4.300E+010	33.541	33.4890	5.750E+010	20.982	20.9544
1.165E+011	32.353	32.3039	1.106E+012	20.869	20.8419
2.845E+011	32.041	31.9924	3.335E+011	20.717	20.6902
4.250E+010	31.408	31.3602	4.090E+011	20.683	20.6566
1.790E+011	31.281	31.2339	2.570E+011	20.515	20.4884
1.170E+011	31.004	30.9575	6.850E+010	20.378	20.3519
1.990E+011	30.700	30.6542	5.665E+011	20.369	20.3427
6.650E+010	30.605	30.5591	8.281E+011	20.328	20.3022
2.400E+010	30.342	30.2966	1.335E+011	20.324	20.2982
5.000E+010	29.951	29.9070	3.350E+010	20.304	20.2780
2.775E+011	29.127	29.0845	3.285E+011	20.248	20.2221
3.250E+010	27.717	27.6767	1.570E+011	20.228	20.2014
1.275E+011	27.340	27.3009	1.485E+011	20.167	20.1407
6.750E+010	27.139	27.1003	2.800E+010	20.141	20.1148
1.424E+012	27.129	27.0899	6.800E+010	20.115	20.0889
5.250E+010	27.034	26.9957	4.068E+012	20.073	20.0469
7.251E+010	26.700	26.6618	1.210E+011	19.782	19.7564
2.175E+011	26.395	26.3578	1.645E+011	19.688	19.6629
1.285E+011	26.383	26.3459	4.750E+010	19.438	19.4135
2.750E+010	25.753	25.7171	2.661E+012	19.349	19.3245
3.650E+010	25.677	25.6410	1.120E+011	19.338	19.3134
1.500E+009	25.503	25.4673	5.460E+011	19.329	19.3041
4.275E+011	24.851	24.8164	5.505E+011	19.250	19.2251
1.450E+010	24.722	24.6880	4.900E+010	19.122	19.0979
2.240E+011	24.489	24.4550	2.300E+010	19.101	19.0769
3.645E+011	24.394	24.3607	4.500E+010	19.048	19.0235
5.050E+010	24.348	24.3150	9.361E+011	18.721	18.6977
1.100E+010	24.094	24.0610	2.150E+010	18.599	18.5755
1.300E+011	24.020	23.9877	6.050E+010	18.458	18.4350
1.760E+011	24.007	23.9738	2.705E+011	18.418	18.3945
6.300E+010	23.987	23.9545	7.451E+010	18.245	18.2220
3.330E+011	23.720	23.6880	3.835E+011	18.056	18.0334
1.515E+011	23.632	23.5999	2.880E+011	18.000	17.9779
4.135E+011	23.316	23.2843	1.000E+009	17.973	17.9506
1.636E+012	23.130	23.0984	1.194E+012	17.936	17.9140
4.350E+010	22.926	22.8951	2.522E+012	17.861	17.8383
2.149E+012	22.841	22.8101	1.068E+012	17.849	17.8273

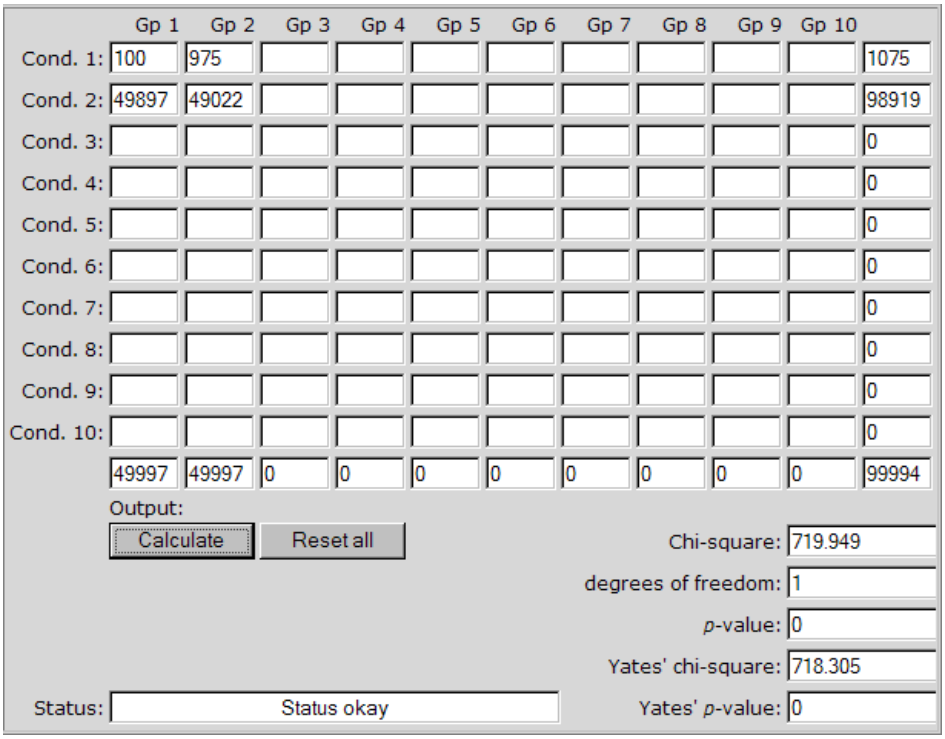


Figure 38: Chi-square with Yates' Correction Linear data

Table 3: Sequential Parallel configured molecule pair Frequency/Power Results

----
Analysis for data file: ParallelColumn01z.txt

Total Number of Coefficients Positive Side of Spectrum
49997
Total Number of Significant Coefficients Positive Side of Spectrum
778
Percentage of Total that are Significant
1.55811
Power in Significant Coefficients div Total Power*100
20.46581
Upper and Lower Frequency Limits Returned by Transform
Lower Frequency Upper Frequency Sampling Frequency
5.000E+008 2.500E+013 5.00000E+013
F2,2R Alpha
6.91 0.001

Top 100 Coefficients Ranked by Magnitude of Power (DC Listed First)					
Hz	Hartley	Power	Hz	Hartley	Power
0.000E+000	n/a	866.6693			
1.500E+009	377.790	371.7146	9.036E+011	22.030	21.8299
5.000E+008	338.651	333.4636	5.850E+010	21.807	21.6086
1.000E+009	287.032	282.9254	9.801E+010	21.638	21.4417
6.000E+009	118.455	117.1527	8.501E+009	21.472	21.2771
5.500E+009	93.668	92.6844	8.801E+010	21.336	21.1423
1.000E+010	80.182	79.3612	4.750E+010	21.258	21.0655
3.000E+009	76.670	75.8910	3.680E+011	20.932	20.7420
9.501E+009	76.575	75.7968	1.375E+011	20.364	20.1800
1.350E+010	66.608	65.9441	4.315E+011	20.346	20.1617
4.500E+009	63.578	62.9485	9.651E+010	20.253	20.0696
1.100E+010	54.500	53.9696	8.531E+011	20.205	20.0222
2.450E+010	52.072	51.5680	2.940E+011	20.037	19.8561
3.850E+010	48.543	48.0766	5.200E+010	19.503	19.3269
5.050E+010	46.730	46.2825	1.605E+011	19.333	19.1580
2.400E+010	44.787	44.3597	1.260E+011	19.310	19.1352
4.250E+010	44.299	43.8768	3.800E+010	19.141	18.9682
9.001E+009	39.279	38.9089	3.500E+009	19.039	18.8674
7.501E+009	38.966	38.5986	7.501E+010	19.033	18.8611
1.250E+010	38.096	37.7376	8.601E+010	18.859	18.6888
4.350E+010	35.982	35.6453	5.150E+010	18.260	18.0956
1.355E+011	33.431	33.1202	1.470E+011	17.822	17.6616
9.201E+010	32.511	32.2091	4.000E+009	17.751	17.5916
5.000E+009	32.255	31.9556	8.231E+011	17.722	17.5628
2.550E+010	31.926	31.6302	6.950E+010	17.714	17.5544
8.901E+010	31.840	31.5441	2.165E+011	17.484	17.3267
1.800E+010	31.166	30.8777	1.400E+010	17.329	17.1730
5.450E+010	31.024	30.7366	1.185E+011	17.200	17.1730
4.900E+010	29.878	29.6020	2.190E+011	17.078	16.9247
6.700E+010	29.743	29.4686	1.100E+011	17.051	16.8979
2.750E+010	29.385	29.1134	2.750E+011	17.000	16.8468
7.000E+009	29.027	28.7590	8.236E+011	16.970	16.8178
7.601E+010	28.871	28.6047	1.415E+011	16.944	16.7922
1.210E+011	28.122	27.8630	5.705E+011	16.931	16.7794
2.000E+009	27.872	27.6158	1.262E+012	16.862	16.7101
1.750E+010	27.097	26.8480	6.635E+011	16.838	16.6868
2.650E+011	26.524	26.2810	1.451E+012	16.733	16.5827
1.398E+012	26.288	26.0468	6.900E+010	16.484	16.3357
7.401E+010	25.779	25.5431	3.485E+011	16.473	16.3255
1.050E+010	24.891	24.6634	2.105E+011	16.453	16.3049
1.625E+011	24.264	24.0425	1.027E+012	16.333	16.1869
1.710E+011	24.221	23.9997	1.230E+011	16.188	16.0433
1.240E+011	24.212	23.9912	5.065E+011	16.042	15.8981
2.115E+011	24.128	23.9079	2.685E+011	16.026	15.8820
6.400E+010	23.979	23.7604	6.050E+010	15.993	15.8499
2.625E+011	23.944	23.7257	7.501E+011	15.975	15.8320
4.300E+010	23.464	23.2498	4.550E+010	15.959	15.8155

2.055E+011	23.449	23.2353	1.320E+011	15.949	15.8065
2.950E+010	23.444	23.2301	2.145E+011	15.923	15.7801
4.855E+011	23.219	23.0078	1.275E+011	15.841	15.6989
4.500E+010	22.755	22.5475	1.611E+012	15.727	15.5861

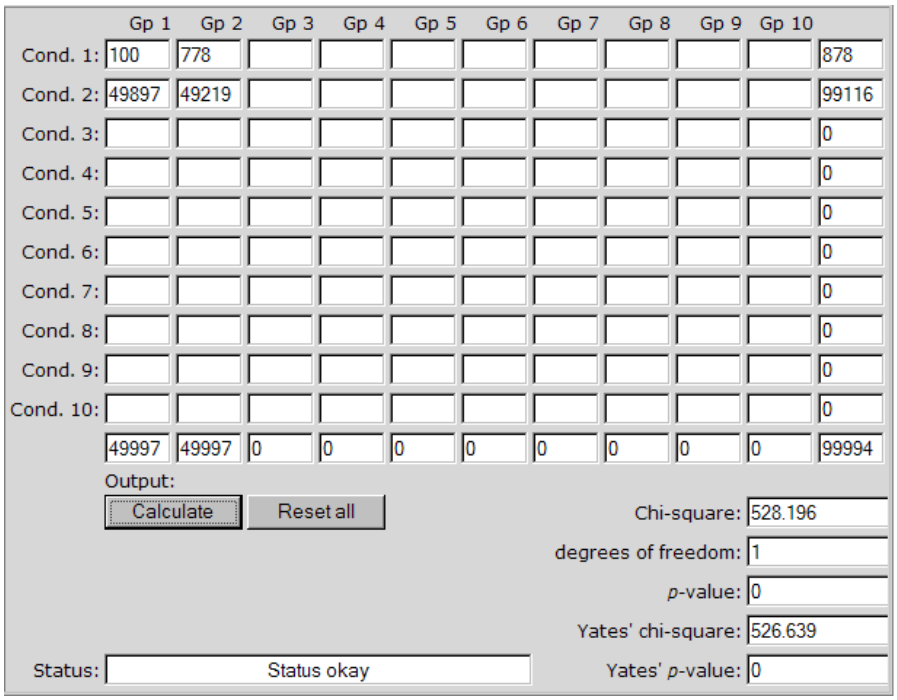


Figure 39: Chi-square with Yates' Correction Parallel data

Table 4: Sequential Perpendicular "T" configured molecule pair Frequency/Power Results

----
Analysis for data file: PerpTColumn01z.txt

Total Number of Coefficients Positive Side of Spectrum
49997
Total Number of Significant Coefficients Positive Side of Spectrum
845
Percentage of Total that are Significant
1.69212
Power in Significant Coefficients div Total Power*100
19.92956
Upper and Lower Frequency Limits Returned by Transform
Lower Frequency Upper Frequency Sampling Frequency

5.000E+008	2.500E+013	5.00000E+013

F2,2R	Alpha
6.91	0.001

Top 100 Coefficients Ranked by Magnitude of Power (DC Listed First)

Hz	Hartley	Power	Hz	Hartley	Power
0.000E+000	n/a	97.5275			
7.501E+009	156.182	155.5469	1.180E+011	21.565	21.5353
2.750E+010	117.770	117.3808	5.500E+010	21.507	21.4774
3.250E+010	90.890	90.6382	1.325E+012	21.440	21.4107
2.000E+010	76.770	76.5791	1.195E+011	21.356	21.3263
6.250E+010	69.631	69.4679	5.350E+010	21.288	21.2591
5.000E+009	66.592	66.4396	1.000E+009	20.812	20.7838
1.750E+010	63.659	63.5176	1.020E+011	20.710	20.6813
4.500E+009	40.266	40.1950	2.300E+011	20.562	20.5344
1.695E+011	39.130	39.0616	2.865E+011	20.424	20.3965
5.000E+010	38.084	38.0182	1.045E+011	20.356	20.3284
1.390E+011	38.056	37.9905	1.710E+011	20.170	20.1426
1.050E+011	36.934	36.8714	4.735E+011	20.138	20.1103
1.820E+011	34.535	34.4785	5.765E+011	20.125	20.0979
8.051E+010	32.859	32.8059	7.361E+011	20.111	20.0841
4.150E+010	32.254	32.2026	2.610E+011	20.077	20.0498
1.095E+011	30.843	30.7948	4.025E+011	19.876	19.8489
2.700E+011	30.113	30.0661	1.268E+012	19.808	19.7815
6.200E+010	29.752	29.7060	4.750E+010	19.677	19.6503
8.001E+009	29.271	29.2257	1.135E+011	19.623	19.5968
5.950E+010	28.962	28.9179	6.150E+010	19.512	19.4854
9.001E+010	28.896	28.8517	1.500E+009	19.495	19.4688
4.525E+011	28.646	28.6024	1.010E+011	19.470	19.4441
5.450E+010	28.554	28.5102	2.800E+011	19.139	19.1130
1.670E+011	27.336	27.2950	1.400E+011	19.134	19.1087
2.015E+011	26.624	26.5839	3.500E+010	19.096	19.0701
1.510E+011	26.588	26.5484	1.540E+012	19.039	19.0131
9.651E+010	26.380	26.3412	1.110E+011	18.915	18.8895
3.470E+011	26.136	26.0978	3.000E+009	18.774	18.7487
3.850E+010	25.598	25.5605	3.150E+010	18.736	18.7109
2.050E+010	24.982	24.9454	2.035E+011	18.723	18.6985
4.850E+010	24.924	24.8881	7.941E+011	18.667	18.6421
4.955E+011	24.346	24.3112	3.000E+010	18.515	18.4901
1.125E+011	24.104	24.0698	1.430E+011	18.430	18.4051
6.500E+009	23.655	23.6208	4.530E+011	18.251	18.2266
6.565E+011	23.575	23.5415	1.740E+011	18.207	18.1827
6.600E+010	23.317	23.2836	6.770E+011	17.861	17.8373
2.000E+011	23.293	23.2602	5.300E+010	17.531	17.5085
1.410E+011	23.277	23.2436	6.525E+011	17.504	17.4810
6.000E+010	23.190	23.1571	4.940E+011	17.358	17.3354
2.585E+011	23.174	23.1407	1.456E+012	17.268	17.2454
2.650E+010	22.805	22.7723	2.564E+012	17.256	17.2338
5.630E+011	22.679	22.6473	7.166E+011	17.206	17.1840
1.250E+010	22.492	22.4604	2.965E+011	17.094	17.0719
3.550E+010	22.408	22.3764	6.300E+011	17.063	17.0408
2.150E+010	22.315	22.2841	2.190E+011	17.060	17.0375
2.500E+009	22.265	22.2342	8.751E+010	17.029	17.0071
3.205E+011	21.952	21.9213	3.000E+011	16.957	16.9346
2.100E+010	21.857	21.8264	7.401E+010	16.904	16.8817
3.885E+011	21.638	21.6079	1.120E+012	16.842	16.8207
27.576E+011	21.603	21.5728	2.350E+010	16.772	16.7501

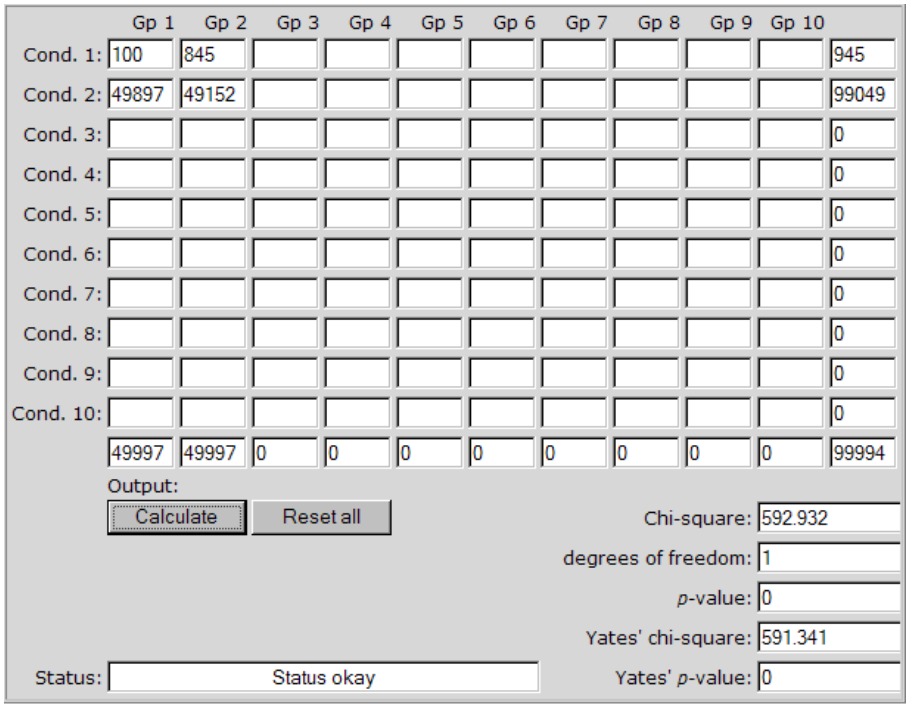


Figure 40: Chi-square with Yates' Correction PerpT data

Table 5: Sequential Skew configured molecule pair Frequency/Power Results

----
Analysis for data file: SkewColumn01z.txt

Total Number of Coefficients Positive Side of Spectrum
49997
Total Number of Significant Coefficients Positive Side of Spectrum
927
Percentage of Total that are Significant
1.85613
Power in Significant Coefficients div Total Power*100
21.55066
Upper and Lower Frequency Limits Returned by Transform
Lower Frequency Upper Frequency Sampling Frequency
5.000E+008 2.500E+013 5.00000E+013
F2,2R Alpha
6.91 0.001
Top 100 Coefficients Ranked by Magnitude of Power (DC Listed First)
Hz Hartley Power Hz Hartley Power
0.000E+000 n/a 4412.5841

2.000E+010	127.294	121.3705	2.187E+012	21.951	20.9736
1.500E+009	118.071	112.5971	2.100E+010	21.861	20.8873
2.700E+010	82.785	79.0026	2.400E+010	21.818	20.8463
5.000E+008	59.338	56.6532	1.200E+010	21.697	20.7306
6.450E+010	58.686	56.0316	4.035E+011	21.632	20.6689
2.545E+011	57.091	54.5101	7.801E+011	21.440	20.4852
5.000E+009	52.179	49.8250	7.391E+011	21.356	20.4056
8.501E+009	43.926	41.9512	2.321E+012	21.344	20.3937
1.745E+011	42.364	40.4615	1.350E+010	21.099	20.1599
5.000E+010	40.342	38.5317	1.340E+012	20.659	19.7392
1.500E+010	40.024	38.2277	8.351E+010	20.643	19.7240
1.050E+010	38.432	36.7088	3.868E+012	20.447	19.5367
1.250E+011	34.647	33.0961	1.194E+012	20.396	19.4883
9.501E+009	34.290	32.7551	1.805E+011	20.019	19.1282
2.600E+010	33.968	32.4474	3.800E+010	19.918	19.0323
3.000E+009	33.512	32.0119	4.700E+010	19.769	18.8894
2.500E+009	33.391	31.8973	2.705E+011	19.741	18.8626
1.000E+009	32.783	31.3162	3.845E+011	19.645	18.7712
4.750E+010	31.689	30.2717	3.150E+010	19.600	18.7279
1.515E+011	30.985	29.6002	2.371E+012	19.583	18.7117
3.695E+011	30.257	28.9046	7.856E+011	19.528	18.6592
6.500E+009	30.241	28.8897	2.320E+011	19.400	18.5368
2.220E+011	28.673	27.3923	4.925E+011	19.260	18.4029
7.446E+011	27.076	25.8676	4.430E+011	19.196	18.3424
2.900E+011	25.772	24.6220	3.200E+010	18.967	18.1233
2.360E+011	25.749	24.6010	2.485E+011	18.853	18.0149
1.000E+010	25.590	24.4485	1.800E+011	18.845	18.0073
1.360E+011	25.409	24.2762	1.175E+011	18.771	17.9364
2.860E+011	25.250	24.1242	3.595E+011	18.526	17.7023
6.330E+011	25.247	24.1211	7.266E+011	18.490	17.6679
6.800E+010	24.808	23.7019	1.850E+011	18.443	17.6227
3.875E+011	24.697	23.5957	1.520E+011	18.401	17.5826
1.900E+011	24.555	23.4600	1.222E+012	18.252	17.4404
8.501E+010	24.503	23.4108	9.251E+010	18.198	17.3885
7.801E+010	24.397	23.3096	3.470E+011	18.068	17.2647
7.651E+010	24.182	23.1042	1.850E+010	17.987	17.1877
4.100E+010	23.910	22.8442	2.030E+012	17.935	17.1380
2.060E+011	23.888	22.8236	5.050E+010	17.846	17.0532
1.345E+011	23.710	22.6534	2.862E+012	17.825	17.0329
2.000E+009	23.502	22.4545	2.508E+012	17.810	17.0179
1.750E+010	23.387	22.3449	7.831E+011	17.775	16.9845
1.415E+011	23.146	22.1150	1.740E+011	17.699	16.9122
2.550E+010	23.137	22.1065	1.723E+012	17.576	16.7944
3.315E+011	23.115	22.0854	1.024E+012	17.574	16.7926
8.776E+011	22.956	21.9334	2.967E+012	17.524	16.7453
7.501E+009	22.688	21.6776	3.102E+012	17.443	16.6679
1.185E+011	22.654	21.6447	3.600E+010	17.405	16.6312
2.130E+011	22.498	21.4962	1.565E+011	17.375	16.6024
8.611E+011	22.374	21.3773	1.099E+012	17.339	16.5687
4.390E+011	22.003	21.0231	8.651E+010	17.333	16.5629

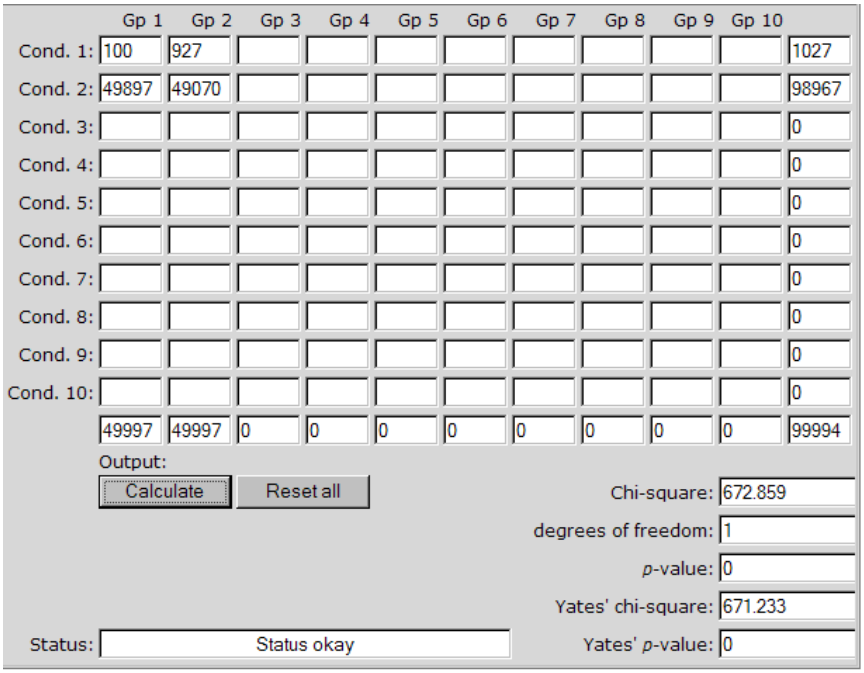


Figure 41: Chi-square with Yates' Correction Skew data

The contingency tables are summarized as below:

Table 6: Contingency Table Analysis Summary

System	Chi-square	Yates' Chi-square	Yates' p-value
Linear	719.949	718.305	< 0.0001
Parallel	528.196	526.639	< 0.0001
Perpendicular T	592.932	591.341	< 0.0001
Skew	672.859	671.233	< 0.0001

In every case the low p-value indicates the association between rows (System) and columns (coefficients significant or insignificant) is very statistically significant and unlikely to be the result of random chance.

General Observations Regarding this Data

There are several notable points to be made about the spectral data thus far. First is the extreme variability between systems of all major indicators. The DC component (portion of the real-time signal that does not average to zero), frequency values and frequency magnitudes all differ widely. A little speculation is needed at this point to pave the way for several conclusions in the next chapter and to provide a transition into the discussion of the last system. Perhaps a little more insight can be gained by extracting a few relevant data points from the large data tables just presented. The percentage of power in significant coefficients, percentage of coefficients that are significant and the top 5 frequencies and their magnitudes for each of the 4 systems have been concentrated into the following 2 tables.

Table 7: Percentage Comparison of 4 Syste

System Name	Percent Power in Significant Coefficients	Percent Significant Coefficients
End-to-End (Linear)	22.77335	1.95214
Parallel	20.46581	1.55811
Perpendicular "T"	19.92956	1.69212
Skew	21.55066	1.85613

Table 8: Percentage Comparison of 4 Syste

Variable	Linear	Parallel	PerpT	Skew
Frequency1	1.00E+10	1.50E+09	7.50E+09	2.00E+10
Frequency2	5.00E+09	5.00E+08	2.75E+10	1.50E+09
Frequency3	1.25E+10	1.00E+09	3.25E+10	2.70E+10
Frequency4	6.25E+10	6.00E+09	2.00E+10	5.00E+08
Frequency5	2.10E+10	5.50E+09	6.25E+10	6.45E+10
Magnitude1	201.1453	371.7146	155.5469	121.3705
Magnitude2	56.6752	333.4636	117.3808	112.5971
Magnitude3	54.3712	282.9254	90.6382	79.0026
Magnitude4	53.926	117.1527	76.5791	56.6532
Magnitude5	50.3925	92.6844	69.4679	56.0316
DC Magnitude	90.5712	866.6693	97.5275	4412.5841
Atom Count	16368	15758	30176	42506

Interestingly, note that in the End-to-End Linear configuration 1.9% of the coefficients account for 22.8% of the power yet in the Parallel configuration only 1.6% of the coefficients account for 20% of the power. Furthermore, note that the most powerful single frequency in the Parallel configuration is about 4 times larger than the most powerful frequency in the Skew configuration. Lastly, take note of the 50 to1 ratio between the DC component of the Skew configuration and the DC component of the End-to-End Linear configuration. Can these large divergences be considered statistical outliers and simply thrown out or are they singularities representing phenomena of the utmost importance? Remember there are only 2 differences between each system, the geometric orientation of the DNA and the size and dimensions (atom count) of each solvent box.

Consider first the bulk distribution in Table 7. Remembering that the water itself is vibrating with great energy the analogy of “water in a bathtub” comes to mind. The configurable parameter in the NAMD simulator known as “Periodic Boundary Conditions” acts in a similar

fashion to a bathtub inasmuch as it serves to contain the water within the system during molecular dynamics. As stated earlier, Periodic Boundary Conditions establish mirror images of the system on all 6 sides to establish the boundaries. It is reasonable to assume that molecular pressure propagating through the boundary could "reflect" back into the system or enter the system from a periodic image much like waves slosh around in a bathtub. This could be analogous to the way electromagnetic waves reflect back from boundaries between materials with dissimilar impedance (EM reflections from mismatched antennas). It is also reasonable to assume that waves may "reflect" off of the DNA molecules themselves further complicating the already "choppy water" wave system. If the simulated molecular system is behaving in a like fashion a large change in "bathtub" dimensions could easily account for the observed changes in bulk distribution of both frequency *and* magnitude of the 'sloshing'.

Next consider the nearly 4 to 1 variability in the top coefficients of Skew and Parallel configurations. At first glance, the size of the Skew configuration being nearly 3 times larger than the Parallel configuration is attractive as a potential correlation variable but becomes inconsistent when all 4 systems are considered. Absent any other obvious contributors the DNA becomes the most interesting prospect. This apparent 'outlier' may be a sign of the very process we are looking for.

The DC magnitude variability exhibits the most pronounced contrast of all. Like the top coefficient variability, the DC magnitude variability also appears at first to be related to system size but again is inconsistent across all 4 systems. Lets explore for a minute what the DC (analogous to Direct Current in an electrical circuit) component represents. From the spectral content tables the lower cutoff frequency of 5.000E8 is listed and is the same for each system. In

practical terms this means that any frequency content contained in the signal below this value would not be *sufficiently* sampled to appear in the spectrum. An important distinction should be made here. It *does not* mean that a frequency below the cutoff does not appear in the data at all because our signal is raw data and has not been filtered at the cutoff point. It means that only a portion of lower frequency signals, if present, will appear in the spectrum. A graphical illustration of an under-sampled signal is shown below in Figure 42.

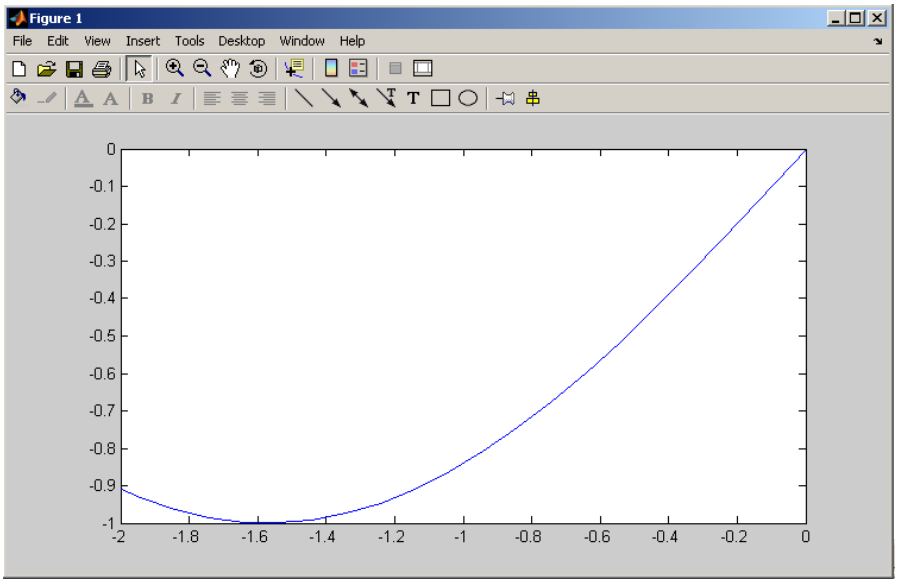


Figure 42: Under Sampled Sin Wave

Figure 42 is a portion of a sin wave with a period of 2π sampled between -2 and 0. An approximate average of this portion of the signal is about -0.5. If this signal were included as part of data transformed into the frequency domain it would manifest itself as a DC component approximately equal to -0.5. Below in Figure 43 we see the same signal sampled for a longer time period of – π to +π representing one complete cycle of the signal.

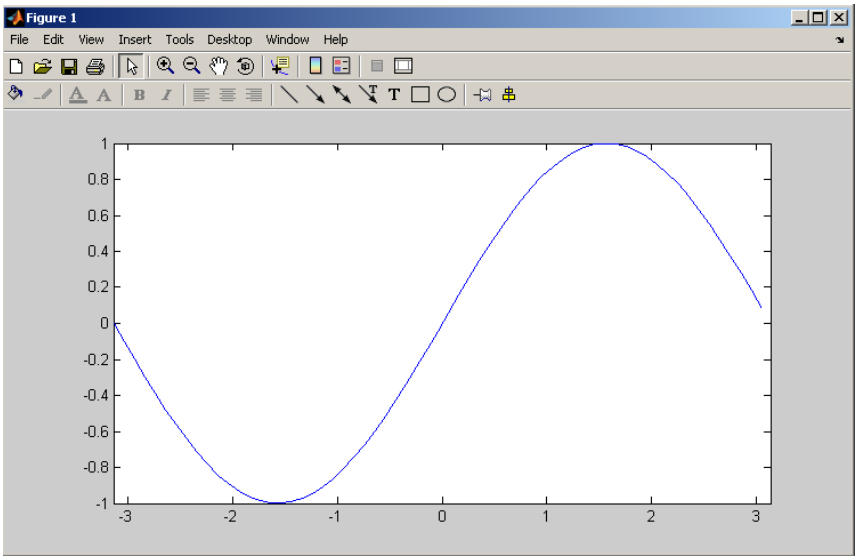


Figure 43: Sufficiently Sampled Signal

An average of this sample set would result in a very different value (zero) than the average of the previous subset.

The implication being that the DC values found in the spectrums of our molecular systems could represent small subsets or sections of large low frequency pressure variations. The concept is demonstrated once again in Figure 44.

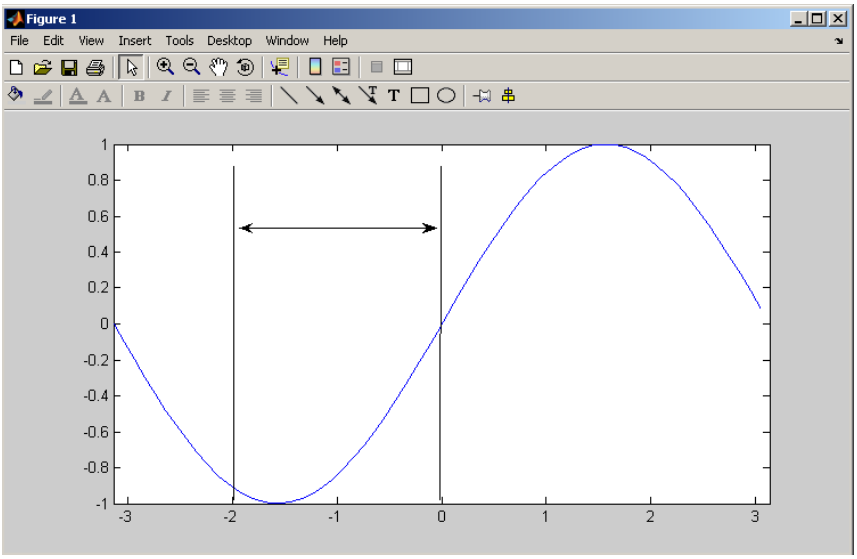


Figure 44: Only a portion of the Signal is Sampled

Although purely speculative now these considerations do provide rationale for the conceptual framework behind our hypotheses and conclusions in later sections.

A New Hypothesis Emerges

At this point abundant frequency content has been identified with the first 4 molecular systems answering several questions and establishing a foundational data set that will undoubtedly prove useful to molecular modelers. There are now of course, many new questions. Is this periodicity an artifact of the models or characteristic behavior of the molecules as we suspect? If the periodicity is characteristic behavior of the molecules, then which molecules? As a direct result of these new questions and a rethinking of $\mathbf{H}_{(\text{Interacting Harmonized Resonance NULL})}$ a new emergent hypothesis has been formulated. Simply stated it is:

Emergent Hypothesis (Sequence Relationship):

$\mathbf{H}_{(\text{Sequence Relationship NULL})}$:= The specific frequencies and amplitudes of the harmonized resonance postulated in $\mathbf{H}_{(\text{Harmonized Resonance NULL})}$ are a direct function of DNA nucleotide sequence. The alternative formulation is:

$\mathbf{H}_{(\text{Sequence Relationship ALT}):}$ = The specific frequencies and amplitudes of the harmonized resonance postulated in $\mathbf{H}_{(\text{Harmonized Resonance NULL})}$ are NOT a direct function of DNA nucleotide sequence.

This hypothesis is not completely new but rather a refinement of $\mathbf{H}_{(\text{Interacting Harmonized Resonance NULL})}$ and as such will be tested for acceptance or rejection *in its place*.

Rationale:

The rationale for this hypothesis is essentially the same as the rationale for the original $\mathbf{H}_{(\text{Interacting Harmonized Resonance NULL})}$ with a narrower scope. It is logical that if DNA specific high power variations do occur the frequency and magnitude would be a function of DNA nucleotide sequence. It is simply an expansion of $\mathbf{H}_{(\text{Interacting Harmonized Resonance NULL})}$ to include another DNA molecule with the additional logic that if DNA specific high power variations do occur the frequency and magnitude would be a function of DNA nucleotide sequence.

As stated in the original rationale, the unique structure of the DNA double helix might broadcast by vibration resonance a frequency and magnitude that are dependant only on the nucleotide base pair sequence. This interacting harmonized resonance might exhibit a magnitude far greater than typical normal mode vibrations most likely at a lower frequency. The intrinsic structural characteristic of each helix strand may determine the frequency, magnitude, and other harmonics similar to how organ pipes determine notes in an organ. This intrinsic property of the structure might cause normally random thermal vibration to synchronize within the hydrophobic

region of the double helix causing segments of the DNA to emit longitudinal pressure waves out into the surrounding water environment. These waves might simply be transverse compressions and rarefactions of the water molecule bonds surrounding the double helix. If a second DNA molecule of similar sequence within close physical proximity were to be exposed to these waves moving through the solvent interaction at a higher level might occur causing further concentration or superposition of vibrational energy into fewer and lower frequencies. This concentration of energy could effectively amplify certain pressure variation frequencies at the expense of others. We are now adding to this rationale the idea that the specific frequencies and their spectral magnitudes are directly related to the nucleotide sequence. Some additional investigating will be required to test this.

Selection of Contrasting System and Sequence

Our emergent hypothesis $\mathbf{H}_{(\text{Sequence Relationship NULL})}$ can be adequately tested if we could somehow establish a relationship between the DNA molecule base pair sequence and the observed frequency content. The intent of the 5th system is to draw a parallel between spectral content and nucleotide sequence potentially substantiating the new hypothesis $\mathbf{H}_{(\text{Sequence Relationship NULL})}$.

Although the linear system was logically the best candidate based on reasoning stated earlier it was decided to use the system exhibiting the most significant evidence of spectral content. Now with a convenient method of quantifying spectral content the most spectrally active configuration could be easily identified and the contrasting sequence could be re-run on only one of the original configurations and accomplish the objective. The data presented in

Table 2 through Table 5 identifies the statistically significant frequency content we are interested in. Using this data we can now individually compare the harmonic content of each molecular system. Because we are in search of a contrast, we want to use the system with the most pronounced frequency characteristics to best highlight system to system differences. It can be reasoned that changing the nucleotide sequence and re-running the system with the most statistically significant coefficients will exhibit the most observable change in coefficients if sequence is actually a factor. With this in mind, the two simplest methods of selecting the best system to re-run are a) the system with the greatest number of significant coefficients as a percentage of total coefficients or b) the system with the greatest ratio of power in significant coefficients to total power. A quick look at Table 7 and Table 8 suggest that re-running the end-to-end linear system, as originally intended, might provide the greatest contrast of sequence related variation.

Dissimilarly Sequenced End-to-End Linear System

Before an antithetical linear system could be constructed an appropriate 'dissimilar' sequence had to be selected. Two factors needed to remain constant between systems in order to maintain the same system electronic charge, the total number of nucleotides and the total of each type of nucleotide. Beyond that, the sequence itself should be totally random. Random number generators are always helpful in these situations because they remove bias. The MATLAB® 'randsample' command provides the exact function needed for this task. For our specific application the syntax for this command was randsample('tataaacgcctataaacgcc',20,false). This command will return a completely random sequence pulled from the pool of nucleotides that

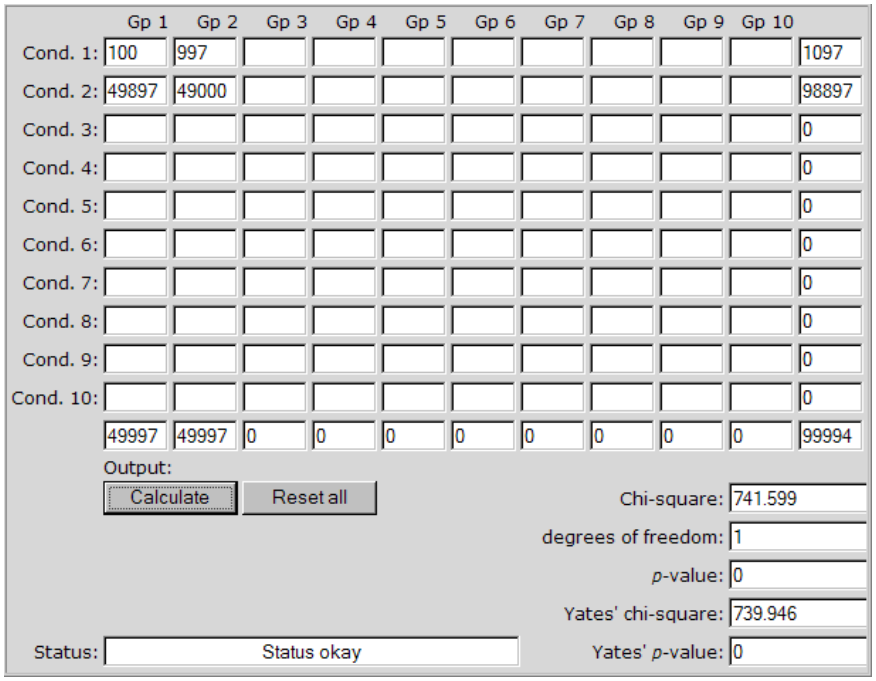
comprise the energetic sequence 'tataaacgcctataaacgcc'. The ‘20’ will return 20 random choices and the ‘false’ will cause the function to choose ‘without replacement’ meaning each time a nucleotide type is chosen it is removed from the available pool thereby eliminating the possibility of changing the overall quantity of each nucleotide. The result is a new sequence consisting of the same number of each nucleotide chosen completely at random. Two consecutive executions of this command returned tgaataacacatctcacacg and atcatatcgcaacagacatc. The linear system was re-constructed, solvated and ionized with these 2 sequences and put through the exact same MD regime as the original linear system. The Ewald pressure calculation run was completed and the entire post simulation data parsing routine was completed. The MATLAB® analysis script was copied to the new Random configuration system directory and the results were generated and tabulated in Table 9.

Table 9: End-to-end linearly configured, Dissimilar Sequenced, Molecule pair Random Frequency/Power

----------					
Analysis for data file: RandomColumn01z.txt					

Total Number of Coefficients					
49997					
Total Number of Significant Coefficients Positive Side of Spectrum					
997					
Percentage of Total that are Significant Positive Side of Spectrum					
1.99614					
Power in Significant Coefficients div Total Power*100					
22.64977					
Upper and Lower Frequency Limits Returned by Transform					
Lower Frequency	Upper Frequency	Sampling Frequency			
5.000E+008	2.500E+013	5.00000E+013			
F2,2R	Alpha				
6.91	0.001				
Top 100 Coefficients Ranked by Magnitude of Power (DC Listed First)					
Hz	Hartley	Power	Hz	Hartley	Power
0.000E+000	n/a	61.2505			
7.501E+009	151.245	150.6996	1.935E+011	21.240	21.2187
2.500E+009	114.448	114.1192	4.165E+011	20.765	20.7445

1.000E+010	93.115	92.8866	1.600E+010	20.639	20.6180
5.000E+009	71.680	71.5354	1.515E+011	20.536	20.5158
8.351E+010	62.254	62.1397	9.801E+010	20.492	20.4715
3.250E+010	52.517	52.4313	5.550E+010	20.183	20.1633
3.000E+009	51.631	51.5475	1.700E+010	20.139	20.1185
7.951E+010	45.934	45.8644	2.105E+011	19.855	19.8355
1.420E+011	45.000	44.9325	2.085E+011	19.807	19.7878
5.750E+010	43.187	43.1238	1.600E+011	19.807	19.7870
3.750E+010	37.366	37.3162	3.055E+011	19.805	19.7855
1.375E+011	36.836	36.7871	1.175E+011	19.661	19.6411
4.250E+010	35.888	35.8413	6.500E+009	19.379	19.3597
5.000E+010	32.249	32.2088	6.820E+011	19.364	19.3446
2.750E+010	32.106	32.0663	3.110E+011	19.339	19.3205
1.795E+011	31.049	31.0110	1.065E+012	19.294	19.2754
1.695E+011	31.033	30.9958	4.605E+011	19.258	19.2389
1.750E+010	31.005	30.9671	4.685E+011	19.206	19.1873
2.520E+011	30.453	30.4168	4.000E+009	19.186	19.1676
3.350E+010	29.909	29.8731	3.815E+011	19.114	19.0952
5.200E+010	29.542	29.5076	5.450E+010	19.106	19.0876
1.235E+011	29.155	29.1212	3.605E+011	19.085	19.0666
1.895E+011	28.633	28.5997	5.500E+010	18.842	18.8240
1.675E+011	28.481	28.4475	4.300E+010	18.816	18.7974
1.885E+011	27.950	27.9181	1.890E+011	18.774	18.7559
5.100E+010	27.908	27.8758	7.636E+011	18.529	18.5108
5.220E+011	27.171	27.1401	1.505E+011	18.507	18.4887
1.000E+009	26.676	26.6459	2.700E+011	18.437	18.4198
1.400E+010	26.602	26.5720	8.551E+011	18.390	18.3724
7.100E+010	25.901	25.8726	6.000E+010	18.386	18.3684
1.095E+011	25.862	25.8331	1.335E+011	18.354	18.3359
4.450E+010	24.953	24.9254	3.475E+011	17.780	17.7635
1.435E+011	24.633	24.6064	4.206E+012	17.741	17.7239
6.475E+011	24.344	24.3175	5.235E+011	17.719	17.7018
7.151E+010	24.263	24.2373	7.221E+011	17.638	17.6213
1.125E+011	23.896	23.8704	1.215E+011	17.633	17.6163
5.845E+011	23.858	23.8324	1.615E+011	17.357	17.3403
9.706E+011	23.750	23.7242	2.481E+012	17.293	17.2766
1.650E+010	23.357	23.3320	1.112E+012	17.269	17.2526
2.250E+010	23.217	23.1921	1.378E+012	17.200	17.1841
4.180E+011	22.930	22.9056	7.501E+010	17.134	17.1182
1.440E+011	22.841	22.8169	1.145E+011	17.130	17.1142
6.855E+011	22.310	22.2872	1.090E+011	17.080	17.0636
6.150E+010	22.236	22.2126	3.105E+011	17.070	17.0540
1.800E+010	22.223	22.2001	2.480E+011	16.991	16.9756
1.089E+012	22.121	22.0983	1.905E+012	16.917	16.9010
9.601E+010	21.958	21.9354	3.024E+012	16.880	16.8648
8.501E+009	21.794	21.7719	1.221E+012	16.608	16.5927
2.788E+012	21.711	21.6888	2.261E+012	16.474	16.4590
5.805E+011	21.583	21.5612	2.145E+011	16.455	16.4401



	Gp 1	Gp 2	Gp 3	Gp 4	Gp 5	Gp 6	Gp 7	Gp 8	Gp 9	Gp 10	
Cond. 1:	100	997									1097
Cond. 2:	49897	49000									98897
Cond. 3:											0
Cond. 4:											0
Cond. 5:											0
Cond. 6:											0
Cond. 7:											0
Cond. 8:											0
Cond. 9:											0
Cond. 10:											0
	49997	49997	0	0	0	0	0	0	0	0	99994

Output:

Calculate | Reset all

Chi-square: 741.599

degrees of freedom: 1

p-value: 0

Yates' chi-square: 739.946

Status: Status okay

Yates' *p*-value: 0

Figure 45: Chi-square with Yates' correction for End-to-End Dissimilar Sequence

Final Spectral Data Results

After all 5 systems were simulated for 2ns the analysis techniques developed above were applied to the trajectories with an alpha of 0.001. The following graphs were designed to illustrate an overall picture of the data from a visual standpoint.

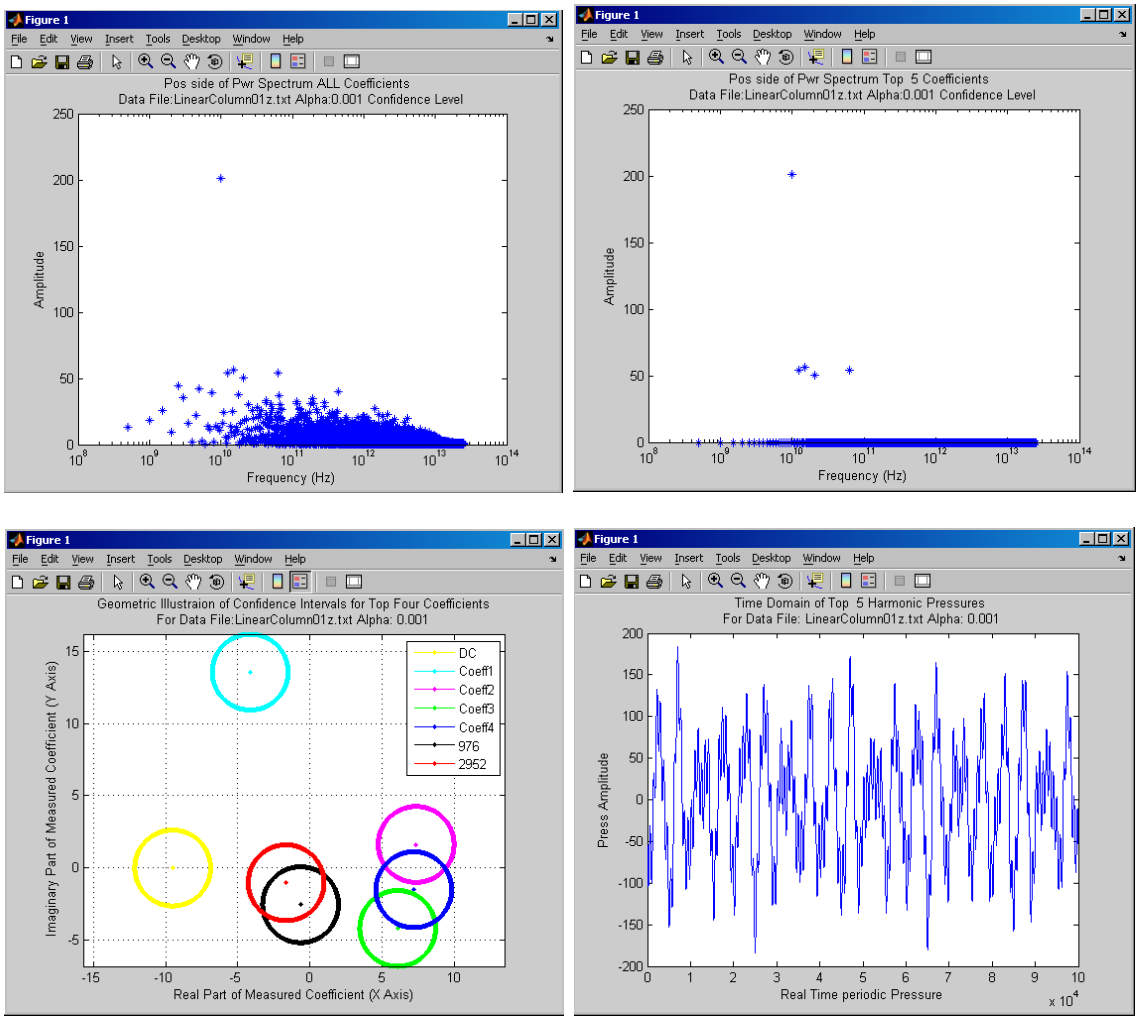


Figure 46: Spectral Content for End-to-end linear Configuration with Identically Sequence Molecules

The upper left graph in Figure 46 illustrates the right side power spectrum amplitude vs. frequency. The upper right graph is the same spectrum with all but the top 5 (ranked by magnitude) coefficients set equal to zero. The lower left graph is the geometric application of a 99.9% confidence interval to the top 4 significant coefficients and 2 insignificant coefficients as well. The lower right graph represents what the top 5 coefficients look like when transformed back into the time domain. Of special note in the confidence interval graph is the inclusion of coefficients 976 and 2952. These coefficients were intentionally included in the program as a

visual “double check” of significance. Coefficient 976, the black data point surrounded by the black circle, is the last coefficient in the spectrum ranked by magnitude to test significant. Coefficient 2952, the red data point surrounded by the red circle, represents a somewhat arbitrary choice intended only to be clearly insignificant and was chosen as the number of the last harmonic times 2 plus 1000 in order ranked by magnitude. The intent of these last to coefficients is simply to test and illustrate graphically the meaning of the confidence interval. Note in Figure 47 below, a zoomed-in view of the last 2 coefficients shows the black circle nearly intersecting the origin, the red circle completely encompassing the origin with a healthy margin, and all other circles notably distant from the origin. These last 2 coefficients are included for reference in all remaining confidence interval graphs.

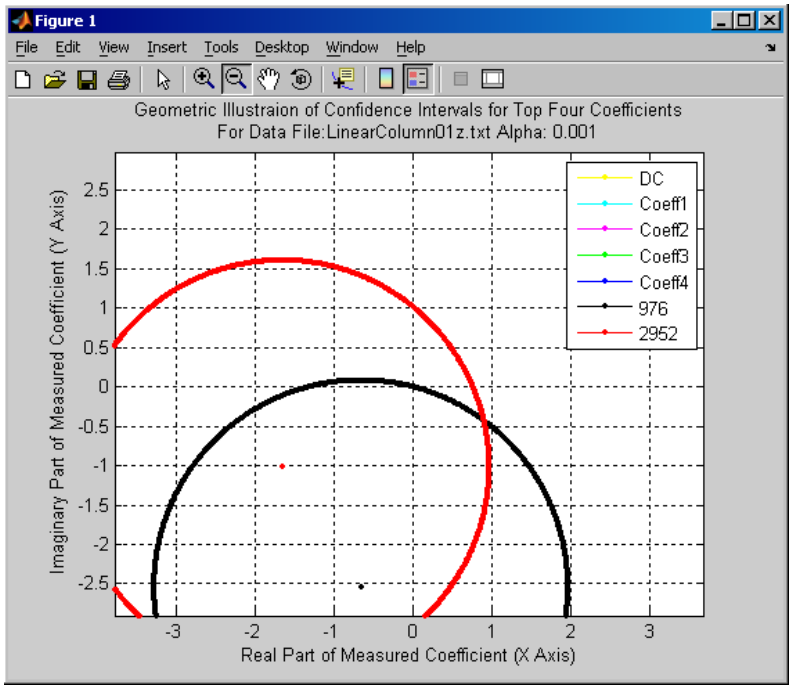


Figure 47: Zoomed In Illustration of Insignificant Coefficients

The time domain graph is intended to convey the physical meaning of the top 5 coefficients. By transforming only the top 5 coefficients back into the time domain we have a

visual representation of actual real time *pressure variation* occurring within the midpoint slab between the DNA. It shows us a digitally filtered image of the original pressure data in real time. Of special note regarding this graph is that the observed behavior is occurring in a cross-sectional slab of the molecular system that contains nothing but water molecules. Following are the graphical results for the remaining 4 systems.

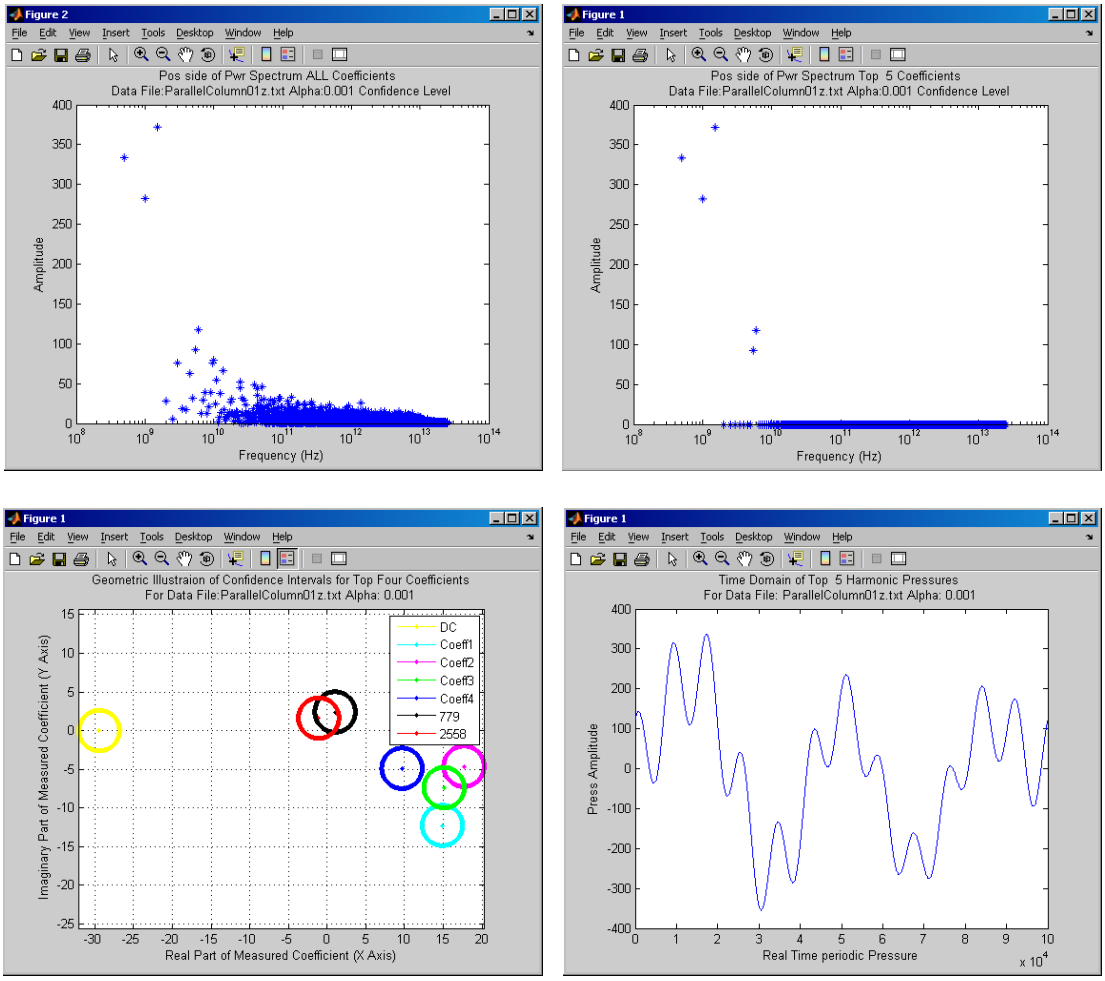


Figure 48: Parallel Configuration Spectral Content

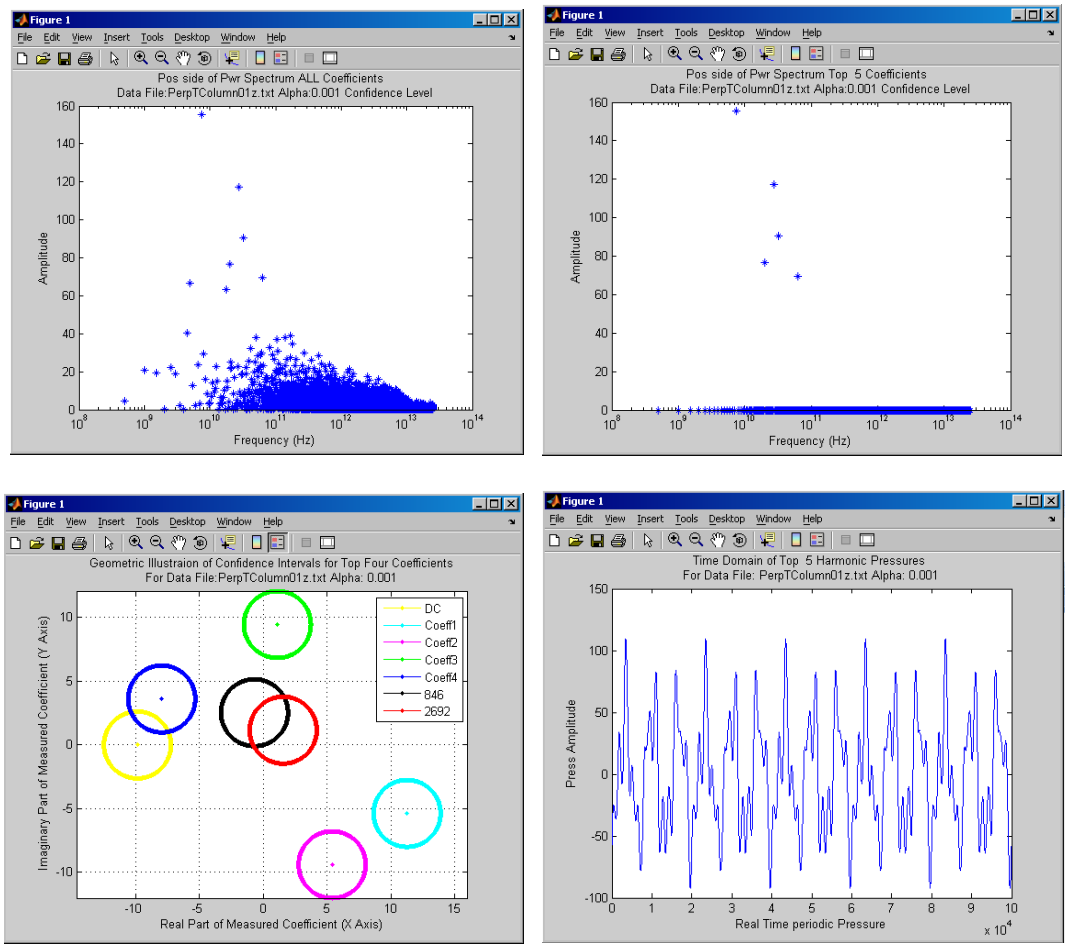


Figure 49: Perpendicular T Configuration Spectral Content

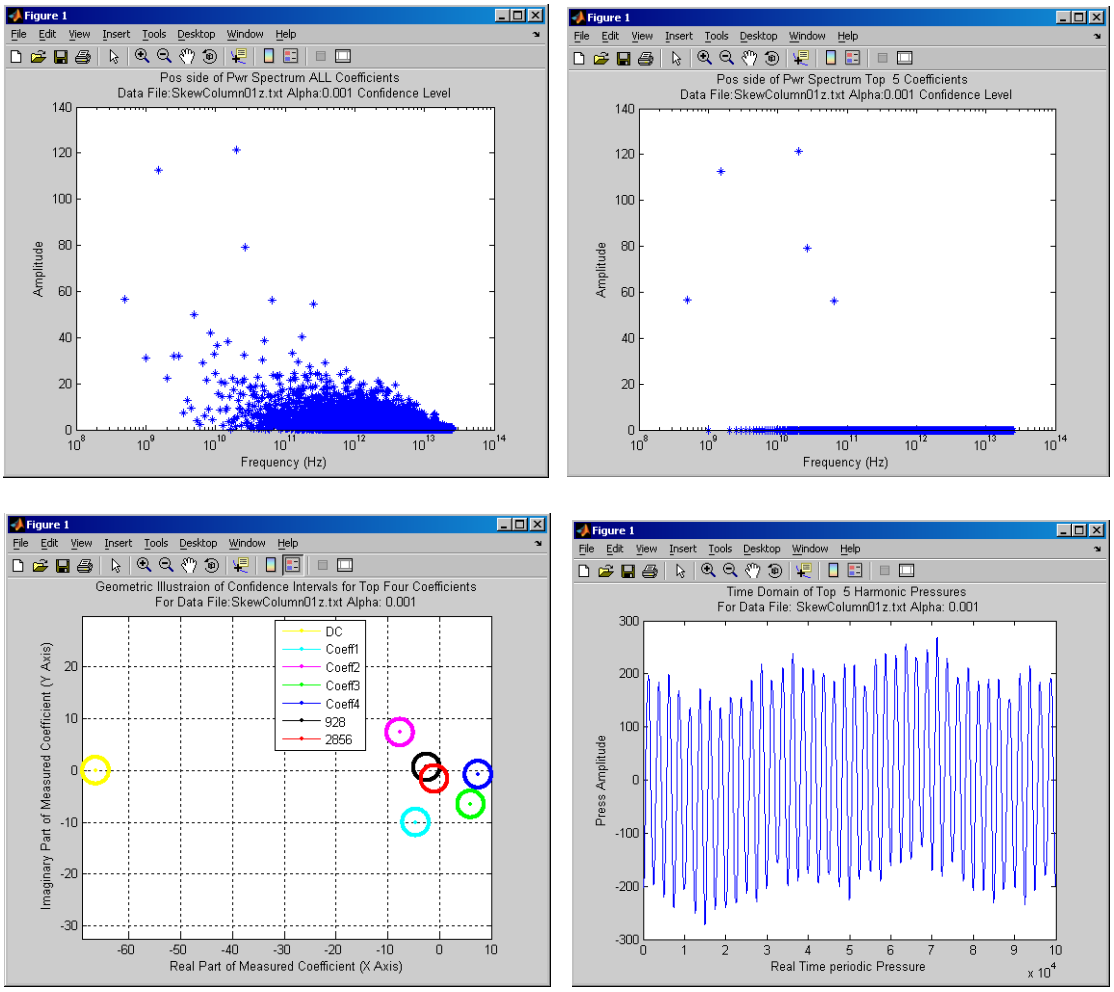


Figure 50: Skew Configuration Spectral Content

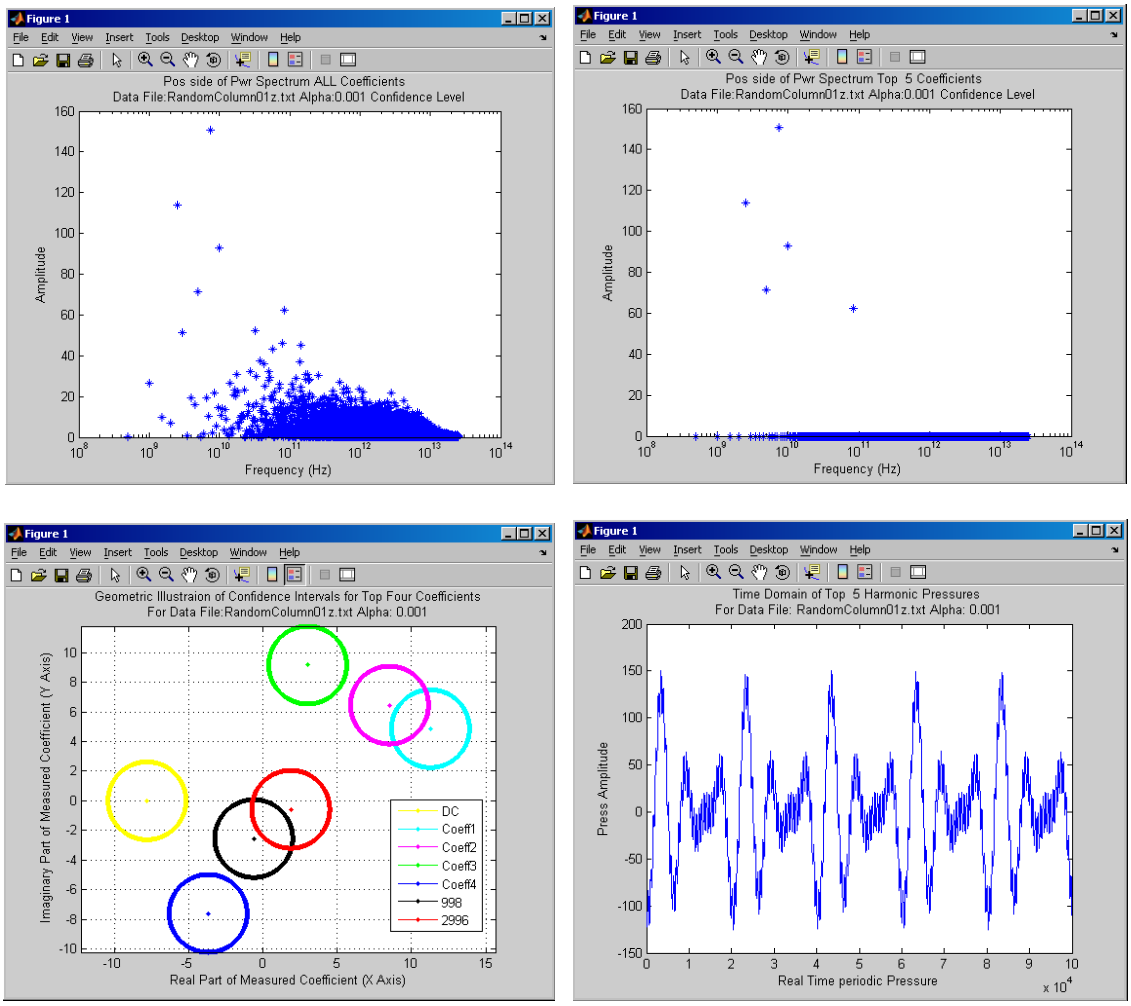


Figure 51: Spectral Content for End-to-end Linear Configuration with Randomly Sequence Molecules

The general observation can be made that the frequency and magnitude vary greatly from system to system.

With data from all 5 systems now available a few key observations can now be made. From Chapter 3 we remember our expectations for this experiment were one or more peaks in the frequency spectrum of the data below the vibrational frequencies of water in the similar sequence system and a relatively flat spectrum taken from the solvent between molecules with

dissimilar sequences. The expected result of a flat spectrum is clearly not the case. In fact visual observation reveals the 2 power spectrum distributions to be quite similar. However, the data does support the expectation of peaks in the frequency spectrum below the vibrational frequencies of water. Although a flat spectrum between dissimilar sequences of DNA would have been a welcome endorsement for $\mathbf{H}_{(\text{Harmonized Resonance NULL}).}$, its absence in no way contradicts this hypothesis providing no firm basis to reject it.

Still, the question remains as to where the identified periodicity is coming from. One could speculate that it is coming from previously mentioned normal mode vibration affecting the slab pressure. If so then why is there so much variation in frequency and magnitude from system to system and why so far below the theoretical range for water normal modes of 10E12 to 10E14? One would expect normal mode vibrations to be consistent between the first 4 systems with variation between the first 4 and the 5^{th}. Again, this is not the case. Add to all this the suspected source of the large spectral DC values being high power low frequency(below Nyquist) pressure variations and you have a pretty good case in support of $\mathbf{H}_{(\text{Harmonized Resonance NULL}).}$

At this point in the research 3 more significant questions naturally arise; "What is the source of the periodicity?", "Why does it vary from system to system?" and "Why is DC so far from the other frequencies for the Parallel and Skew systems?"

Programming Verification

This surprisingly large amount of variable periodicity will naturally cause a skeptical investigator to question the MD simulation algorithms themselves and one's methods. This

finding is an opportunity to temporarily divert the discussion to programming verification. With such a vast array of programming tools custom crafted for this endeavor the inevitable question pops up "Could there be a bug in the software?". The operation of the existing NAMD simulator and the VMD analysis tools are expected to be accurate based on what is now long-term development and a widespread user community providing near real-time feedback to the developers. Verification of the custom PERL and MATLAB® code is a more pressing challenge. Verification of every line of code for this research is not feasible but 2 tests were devised that lend significant confidence to the quality of the analysis.

The first test was primarily intended to verify one of the most difficult data manipulations of the research, the accurate scaling of the X-axis on the Fourier Spectrum in MATLAB® plots. For reasons beyond the scope of this research MATLAB® does not scale the X-axis when it returns the transformed data, the user must do this. It was decided the best way to test the accuracy of the X-axis scale as well as the entire process from reading data to outputting spectrum graphs was to insert a test signal into real data and run it through the analysis. If the analysis was correct the inserted signal would show up in the spectrum exactly where expected verifying the entire chain of calculations. A script was written titled "Install_test_signal_to_orig_data_xyz.m" (included in the Appendix) and executed on real data from the linear system. The script opened up the actual summed pressures and added in an easily recognizable frequency (3.333E11) at a large magnitude (5E2) to act as a flag. The summation was then written out to a different file in the same format. The normal production script was then used to process the modified data file resulting in the following spectrum;

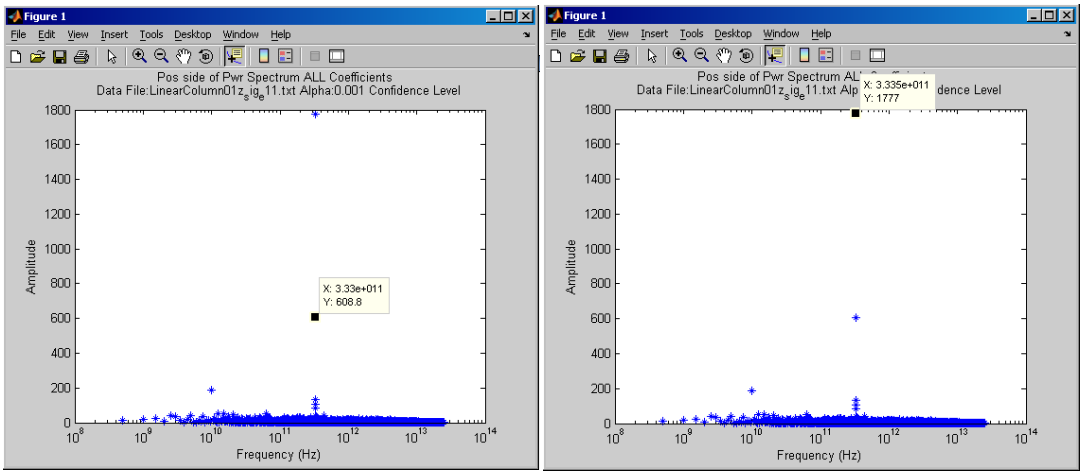


Figure 52: Linear Spectrum with 3.333E11 Test Signal Installed

Note the 2 data pointes highlighted by the data cursor clearly report very significant frequency content at the precise location of the installed signal. The existence of 2 data points illustrates and reminds us of the discreet nature of the transform.

The second test was devised to help verify the statistical assessments made on the power spectrums. The same methodology was used as with the first test, generate a data set with a known synthetic component and run it through the regular production analysis programs. In this case the linear data set was loaded into MATLAB® and the mean and standard deviation were calculated with the appropriate commands. Then, a Gaussian data set was generated using that exact mean and standard deviation with these MATLAB® commands:

```
mu=mean(pressure);
sigma=std(pressure);
R=normrnd(mu,sigma,1,n);
save ./LinearColumn01z_synthetic.txt -ascii R
```

These commands resulted in a Gaussian data set with mean and standard deviation known to be equal to an actual data set stored in the same format as the raw pressure data. This file was then processed with the production analysis program to produce the results below:

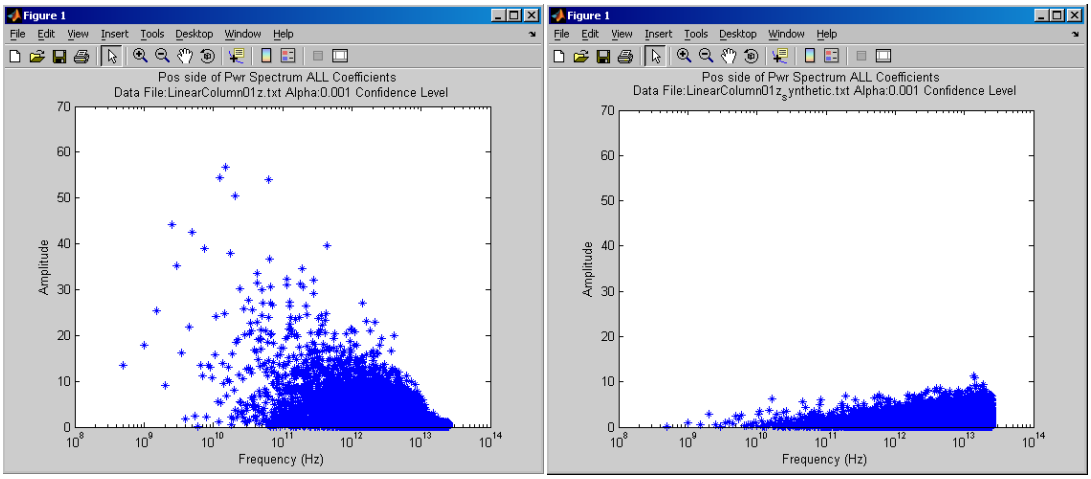


Figure 53: Linear System Spectrum vs. Gaussian Synthetic Data

In the spectrums in Figure 53 the y axes have been set to a value of 70 for comparison purposes. The non-Gaussian nature of the Linear system pressure data becomes visually evident corroborating the statistical assessment. Of particular interest however is the summary printout for the Gaussian data. Please note the 2 highlighted measures in the header section of the summary file excerpt in Table 10 below:

Table 10: Excerpt of Gaussian Data Results

Analysis for data file: LinearColumn01z_synthetic.txt

Total Number of Coefficients
49997
Total Number of Significant Coefficients Positive Side of Spectrum
52
Percentage of Total that are Significant Positive Side of Spectrum
0.10601
Power in Significant Coefficients div Total Power*100
0.82430
Upper and Lower Frequency Limits Returned by Transform
Lower Frequency Upper Frequency Sampling Frequency
5.000E+008 2.500E+013 5.00000E+013
F2,2R Alpha
6.91 0.001

Top 100 Coefficients Ranked by Magnitude					
Hz	Hartley	Power	Hz	Hartley	Power
0.000E+000	n/a	80.0259			
1.373E+013	11.351	11.3395	1.637E+012	6.944	6.9373
1.396E+013	10.937	10.9265	1.975E+013	6.942	6.9352
1.897E+013	9.482	9.4727	1.698E+013	6.940	6.9335
1.407E+013	9.220	9.2113	2.170E+013	6.897	6.8904
1.505E+013	9.149	9.1405	2.294E+013	6.892	6.8855
1.798E+013	8.921	8.9124	2.489E+013	6.884	6.8776
1.192E+013	8.654	8.6457	2.250E+013	6.883	6.8762
3.908E+012	8.550	8.5422	1.945E+011	6.867	6.8607
4.505E+012	8.524	8.5158	1.496E+013	6.845	6.8383
1.470E+013	8.494	8.4858	1.700E+013	6.825	6.8188
1.913E+013	8.355	8.3475	2.310E+013	6.822	6.8162
8.273E+012	8.208	8.2001	1.893E+012	6.812	6.8061
1.115E+013	8.179	8.1718	1.032E+013	6.809	6.8028
2.988E+012	8.136	8.1279	2.335E+013	6.777	6.7712
1.111E+013	8.044	8.0360	1.557E+013	6.773	6.7668
3.745E+012	8.034	8.0266	9.325E+012	6.772	6.7656
6.118E+012	7.978	7.9708	1.515E+013	6.751	6.7452
1.691E+013	7.943	7.9357	1.090E+013	6.749	6.7432
2.057E+013	7.941	7.9333	2.234E+012	6.717	6.7108
4.302E+012	7.764	7.7571	1.463E+013	6.705	6.6992
2.827E+012	7.755	7.7473	7.964E+012	6.692	6.6857
1.657E+013	7.649	7.6417	4.853E+012	6.683	6.6770
5.321E+012	7.523	7.5156	2.339E+013	6.670	6.6643
5.473E+012	7.513	7.5064	5.737E+012	6.667	6.6608
9.222E+012	7.508	7.5011	1.855E+013	6.648	6.6424
1.283E+013	7.441	7.4345	1.005E+013	6.643	6.6372
8.080E+012	7.419	7.4126	1.281E+013	6.638	6.6320
1.719E+013	7.371	7.3639	1.911E+013	6.618	6.6115
2.734E+012	7.346	7.3389	5.259E+012	6.610	6.6044
7.956E+012	7.324	7.3170	5.435E+011	6.590	6.5838
7.307E+012	7.317	7.3100	1.194E+013	6.589	6.5831
8.964E+012	7.301	7.2939	1.728E+013	6.580	6.5745
7.222E+012	7.272	7.2653	2.163E+012	6.558	6.5524
2.256E+013	7.258	7.2512	8.919E+012	6.539	6.5335
8.564E+012	7.240	7.2329	2.089E+013	6.537	6.5314
8.796E+012	7.223	7.2161	4.415E+012	6.531	6.5247
1.382E+013	7.172	7.1650	2.236E+013	6.517	6.5115
1.207E+013	7.164	7.1571	1.818E+013	6.513	6.5075
1.515E+012	7.126	7.1198	1.834E+013	6.503	6.4970
2.169E+013	7.115	7.1084	6.590E+011	6.487	6.4810
2.392E+012	7.113	7.1063	5.794E+012	6.483	6.4769
1.140E+013	7.110	7.1038	5.196E+012	6.478	6.4718
2.389E+013	7.108	7.1010	2.413E+013	6.476	6.4706
1.337E+013	7.072	7.0657	1.209E+013	6.454	6.4483
5.884E+012	7.058	7.0517	1.868E+012	6.435	6.4291
7.721E+012	7.045	7.0383	5.015E+011	6.422	6.4164
1.675E+013	7.043	7.0365	1.843E+013	6.376	6.3704

5.885E+011	7.006	6.9992	2.401E+012	6.358	6.3526
2.124E+013	6.950	6.9432	4.960E+012	6.355	6.3496
7.677E+012	6.945	6.9384	1.201E+013	6.347	6.3409

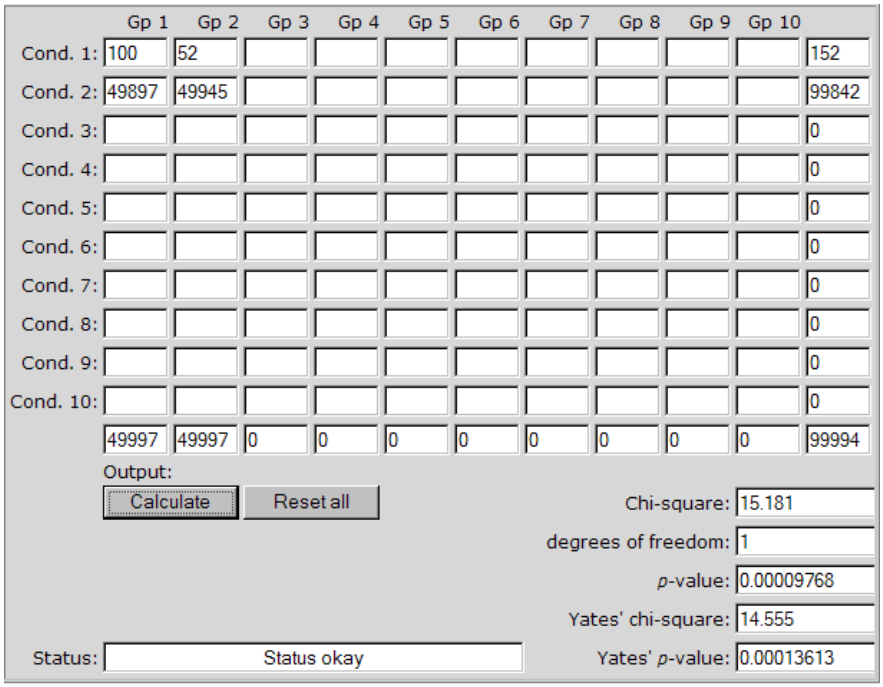


Figure 54: Chi-square with Yates' Correction on Gaussian Data

At first glance it may seem odd that 0.10601% (52) of the Fourier Coefficients returned by the transform test statistically significant from a data set that is known to be Gaussian. It only seems odd until we recall our chosen alpha of 0.001. From the derivation of our test we recall that an alpha of 0.001 tells us that there is a less than 0.1% chance of being wrong. Knowing the data set is Gaussian, the results are observably wrong concerning 52 of the coefficients. This equates to 52 out of 49997 total coefficients or 0.104% errors, almost exactly as expected. Conversely, the Summary printout for the sequential linear system data shows

1.95214% of the total that test significant, well above the 0.1% threshold. The linear summary header is reproduced again below for convenient reference.

Table 11: Sequenced Linear Summary Header Reproduced

Analysis for data file: LinearColumn01z.txt

Total Number of Coefficients Positive Side of Spectrum
49997
Total Number of Significant Coefficients Positive Side of Spectrum
975
Percentage of Total that are Significant
1.95214
Power in Significant Coefficients div Total Power*100
22.77335
Upper and Lower Frequency Limits Returned by Transform
Lower Frequency Upper Frequency Sampling Frequency
5.000E+008 2.500E+013 5.00000E+013
F2,2R Alpha
6.91 0.001
Top 100 Coefficients Ranked by Magnitude

The result of these tests affords us considerable confidence in the overall programmatic chain of events and acceptance of $\mathbf{H}_{(\text{Resonance NULL})}$. Furthermore, the bulk of program and script development was done on a template basis that began with the linear system. This template method provided consistency as each subsequent system was developed and processed.

With added confidence in our results so far we can return the discussion to the general research questions at hand.

Searching for Sequence Effects

The final step in the investigation now is to specifically address the emergent hypothesis $\mathbf{H}_{(\text{Sequence Relationship NULL})}$ that the specific frequencies and amplitudes of the harmonized resonance

postulated in $\mathbf{H}_{(\text{Harmonized Resonance NULL})}$ are a direct function of DNA nucleotide sequence. In light of the periodic behavior found in simulated intermolecular pressures it is most desirable to determine if the pressure variation is related to DNA sequence. Additional testing is now needed to determine if the contribution of pressure variation, defined in terms of significant frequencies, is attributed to the two factors that remained constant between systems in order to maintain the same system electronic charge, the total number of nucleotides and the total of each type of nucleotide AND is attributed to the change in DNA sequence. Going beyond the proposed scope of this project in search of any clues as to the origin of the periodic behavior, significant frequency content from all 5 systems was examined for similarities or patterns. Several PERL programs were written (included in the Appendix) to sort the output data from all 5 molecular systems and compile lists of significant frequency matches. The programs were run and the results printed to text files readable by Excel. The files consisting of 995, 778, 845, 927, and 997 significant frequencies observed in the identically sequenced linear, parallel, perpendicular, skew, and dissimilarly sequenced linear data were imported into Excel where the matches found were then grouped by association with the energetic sequence or the random sequence. Surprisingly there were only 28 statistically significant frequencies that appeared significant in the frequency spectrum of all 4 systems that contained the energetic sequence. Hence one can conclude that the vast majority of the significant frequency content is attributed to system configuration, not sequence. When the matching process was run to include the 3 energetic sequence systems and the linear system with the random sequence only 21 of the original 28 frequencies showed up as significant. This means that 7 specific frequencies that were significant in the spectrum when the energetic sequence was present in the molecular system

turned up missing from the significant frequency spectrum when the energetic sequence was missing from the molecular system, inferring that sequence has a role. Also the dissimilarly sequenced linear system also has 14 new significant frequencies, see Table 12, that were not observed in the list of significant frequencies common to the four identically sequenced systems, further inferring that sequence has a role. Although the observed omissions on the spreadsheet certainly support a correlation between sequence and frequency what about the 14 new frequencies that became significant in the spectrum when the random sequences were present? Since the molecules are now completely different it seems unlikely but not conclusive that the 14 new frequencies can be considered the result of harmonized resonance arising from sequence re-ordering in the base pair sequence.

One possible alternative explanation for the differences in the list of significant frequencies comes to mind when we recall a few things about power spectrums. We need to remember we are dealing with only a sample of a continuous signal. The Fast Fourier Transform within MATLAB® relies on the Nyquist theorem that essentially states the frequency (Nyquist) =1/2v where v is the sampling rate. (Moler, 2004) In theory, it simply means we have to sample a signal at twice its frequency to accurately reconstruct it. This premise is what establishes the upper and lower boundaries of the power spectrum generated by MATLAB®. There is a problem with direct application of this theorem to real world data known as *aliasing*. This means that since our spectral analyses are actually incomplete due to sampling limitations there could be frequency content both above and below the upper and lower boundaries of our spectrums and we would not know they are there. Furthermore, if frequency content does exist above our Nyquist limitation it will tend to 'fold' back into the lower frequencies corrupting them. The

higher frequencies could take on the *alias* of lower frequencies that are not really there. Without prior knowledge of exactly what frequency content is actually in the data we can't adequately filter for it and have to assume some aliasing occurs. With this in mind we can consider the spreadsheet results from a different perspective. It can be reasoned that the frequencies in yellow might be systemic in nature (possibly functions of the DNA structure, the simulator algorithms, the models or even normal mode vibration of the water) and exist at a power level sufficient to remain statistically significant regardless of nucleotide sequence. Consistent with the previous statement about the 7 missing frequencies in red, the sequence change from two energetic sequences to two random sequences could easily change the vibrational characteristics of the system enough to suppress or just slightly redistribute some of the power in the spectrum moving the red frequencies below a significant level. This same reasoning can be applied to the green frequencies that only became significant when the random sequence was present in the system. The concept of potential missing effects from the energetic sequence combined with additional effects of the new random sequence might just be illustrating a changing power distribution landscape that causes the observed changes in the specific frequency magnitudes of interest to us.

Regardless, without further data it cannot be determined if the 7 missing and 14 new frequency values are directly related to a specific nucleotide sequence, related to systemic variables or are simple indications of a random change in the power distribution. Until many more iterations of this experiment are run and analyzed there are simply too many alternative explanations for the data to safely assume there is DNA sequence interaction. It is possible that this entire simulation represents only a single anecdotal data point with respect to frequency and

magnitude and so provides no firm basis on which to accept the null. The spreadsheet with the tabulated results color coded for clarity can be found below in Table 12.

Table 12: Spreadsheet Tabulating Frequency Matches (in Hz)

Comparison of Statistically Significant Matching Frequencies	
Grouped by Association with the Energetic Sequence	
tataaacgcctataaacgcc	
and by Association with 2 Random Sequences	
tgaataacacatctcacacg and atcatatcgcaacagacatc	
4 system Z matches	4 system Z matches
With Energetic Sequence	With Random Sequences
1.00E+09	1.00E+09
1.50E+09	1.50E+09
3.00E+09	3.00E+09
4.50E+09	4.50E+09
5.00E+09	5.00E+09
6.50E+09	6.50E+09
7.50E+09	7.50E+09
8.50E+09	8.50E+09
1.05E+10	1.05E+10
1.10E+10	1.10E+10
1.25E+10	1.25E+10
1.75E+10	1.75E+10
2.00E+10	2.00E+10
2.10E+10	2.10E+10
2.15E+10	2.15E+10
3.00E+10	3.00E+10
4.75E+10	4.75E+10
6.25E+10	6.25E+10
1.38E+11	1.38E+11
1.70E+11	1.70E+11
2.19E+11	2.19E+11
6.05E+10	1.20E+10
8.65E+10	2.25E+10
1.21E+11	3.85E+10
1.21E+11	4.10E+10
1.71E+11	6.20E+10
2.63E+11	7.05E+10
4.79E+11	8.05E+10
	1.15E+11
	1.81E+11
	2.22E+11
	2.42E+11
	2.73E+11
	3.52E+11
	5.22E+11

Summarizing the results of Experiment #2 we find the simulations produced substantial evidence in support of $\mathbf{H}_{(\text{Resonance NULL})}$ and $\mathbf{H}_{(\text{Harmonized Resonance NULL}).}$ The simulations did not produce the fundamentally conclusive evidence to support or reject $\mathbf{H}_{(\text{Sequence Relationship NULL})}$. Even though the intriguing periodicity does allow for a relationship between the DNA molecules, their sequence and the frequency/magnitude of periodic pressure variation to exist; such a relationship cannot be conclusively inferred from this data. Although neither the bulk spectral content found nor the frequency variation observed contradict $\mathbf{H}_{(\text{Sequence Relationship NULL})}$ the results remain inconclusive,

These result cry out for further investigation including at a minimum multiple repetitions of the simulations as currently designed and new configurations designed to answer the multitude of questions that have arisen.

CHAPTER 5: CONCLUSIONS

Chapter 5 Abstract:

A brief recap of the importance of DNA research and the benefits of MD simulation as a research tool is presented. The experimental procedures used and their individual results are briefly summarized. A clear visionary direction for future research is laid out along with identification of specific questions that need answers and large gaps in current knowledge that need filling. The experiments testing $H_{(\text{Simulate Observed Closure NULL})}$ and $H_{(\text{Resonance Causes Closure NULL})}$ are discussed along with reasons for rejecting them. The experiments testing $\mathbf{H}_{(\text{Resonance NULL})}$, $\mathbf{H}_{(\text{Harmonized Resonance NULL})}$ and the emergent hypothesis $\mathbf{H}_{(\text{Sequence Relationship NULL})}$ are discussed along with the reasons for accepting $\mathbf{H}_{(\text{Resonance NULL})}$ and $\mathbf{H}_{(\text{Harmonized Resonance NULL})}$. The sequential hypothesis is discussed at length because some of the results seemed to be related to sequence but the relationship could not be proven. Hence the results were declared inconclusive regarding $\mathbf{H}_{(\text{Sequence Relationship NULL})}$. Lessons learned are addressed in final comments concluding that until several key questions are answered and the apparent in-ability of MD simulation to reproduce laboratory proven DNA segregation is resolved, MD modeling should be considered unsuitable for further investigation of Homologous Chromosome Pairing.

Summary

This research paper began with a cursory appraisal of the significance of the DNA molecule. Recent large scale research efforts were discussed along with the growing collection of mysteries that remain unexplained. The introductory chapters developed a sense of the

enormous underlying significance of the DNA molecule and the equally enormous void of knowledge that currently exists about it. It was shown that the integral relationship DNA has to so many un-explained biological phenomena together with the impact those phenomena have on our civilization establishes a premium value on almost any research related to DNA. The mindset of current DNA research was characterized and it was suggested that existing paradigms in biology are ill-equipped to explain most of what we still don't comprehend about the molecule. A ubiquitous yet completely mysterious biological phenomenon, homologous pairing, was singled out as an example of DNA behavior we are at a complete loss to explain.

A paradigm changing idea was hypothesized as a potential explanation for many observable, yet poorly understood DNA phenomenon, especially homologous pairing. The concept of molecular interaction through harmonized vibrations was introduced and explored in detail. It was suggested that this phenomenon, if it existed, would not only help explain homologous pairing but might spill over and influence many basic theories of chemistry and physics as well.

It was declared that introductory research needed to be done and computational molecular simulation was a very cost effective method for doing said research due to the growing availability of high-performance-computing hardware and a rapidly maturing family of molecular simulation software.

A 2-pronged investigation based solely on computational simulation was proposed addressing 2 separate objectives. The first was to effectively simulate DNA segregation like that observed by the Imperial College team and establish a basis for extrapolating the concept to

homologous pairing. The second was to look for, identify and characterize harmonized resonant intermolecular vibration between solvent immersed DNA molecules if it existed.

Conclusions

A detailed research plan was developed to accomplish the investigation with computationally intensive simulations. Experiment #1 was designed to investigate $H_{(\text{Simulate Observed Closure NULL})}$ and $H_{(\text{Resonance Causes Closure NULL}).}$ It produced no supporting evidence for $H_{(\text{Simulate Observed Closure NULL})}$ therefore that hypotheses must necessarily be rejected. Surprisingly, the complete lack of any closure at all produced by the massive simulation presents a real challenge with respect to testing $\mathbf{H}_{(\text{Resonance Causes Closure NULL})}$. Strictly speaking this particular outcome neither supports nor contradicts the hypothesis and so really doesn't apply to a positive or negative finding. Regarding $\mathbf{H}_{(\text{Resonance Causes Closure NULL})}$ the only final conclusion available to us is that we were unable to actually test it.

Experiment #2 designed and later refined to investigate $\mathbf{H}_{(\text{Resonance NULL})}$, $\mathbf{H}_{(\text{Harmonized Resonance NULL})}$ and the emergent hypothesis $\mathbf{H}_{(\text{Sequence Relationship NULL})}$ produced more interesting results. As detailed in Chapter 4 the simulations produced substantial evidence in support of $\mathbf{H}_{(\text{Resonance NULL})}$ without contradictions or complicating factors. Further consideration of the data and subsequent analysis provide sufficient reason to also accept $\mathbf{H}_{(\text{Harmonized Resonance NULL})}$. On the other hand, the simulations did *not* produce fundamentally conclusive evidence in support of $\mathbf{H}_{(\text{Sequence Relationship NULL})}$. As stated in Chapter 4, parametric statistics were applied to the spectral content of Z axis pressure variations in the water between DNA molecules resulting in statistically significant periodic behavior. This intriguing periodicity was not found to be contradictory to a potential relationship between the DNA molecules, their sequence and the frequency/magnitude of

periodic pressure variation. Because of this it is tempting to infer that DNA-DNA interaction is the source of that periodicity but there are just too many other plausible explanations and too many un-answered questions to safely draw that secondary conclusion without more data. In the final analysis the results remain insufficient to accept or reject $\mathbf{H}_{(\text{Sequence Relationship NULL})}$ and given the extent of our experiments can only be considered inconclusive.

Experimental Limitations

There are 2 facets of the experiments that were limited. The first was the physical size of the virtual systems tested. Because the size of each system is directly proportional to the atom count which is proportional to how long it takes the simulator to calculate a single iteration, it is necessary to minimize the size to minimize how long it takes to run the simulations. With MD simulations a small reduction in system size can easily result in days or weeks shorter run times. Therefore all the systems were designed optimally for size and were possibly too small. The second limiting factor is simulation run time. With MD simulations, run times can be extremely long. Again, with experiments that are optimized to complete within a reasonable amount of time, the run time may be too short.

Lessons Learned

If I could repeat this endeavor from the beginning knowing what I know now I would make 2 significant changes. I would allocate twice the run time resources to the Phi X 174 experiment for a minimum of 4us simulation time. The assumption remains that resonant

closure is the result of a self starting stochastic molecular event. Allocating twice the time to the test might produce very different results. The second major change I would make is I would not test multiple geometric configurations. Knowing now the results of the end-to-end linear and the parallel configurations I would allocate all available resources to larger versions of those systems. Larger systems would allow longer run times thus producing much higher value data.

Future Research

In retrospect Experiment #1 produced the most surprising results. It is a clear case of a computational simulation NOT reproducing what we know actually happened in a laboratory environment. Why did segregation known to occur in-vitro during the Imperial College investigation not occur in this simulation? This apparent contradiction needs further investigation to find out why. The logical next steps would be:

- Since the Imperial College experiment ran for 2 weeks it is not feasible to reproduce such a run time with complete fidelity. Still, it would be time well invested to extend the runtime by resuming the Phi X 174 simulation from its current state allowing the simulation to continue until the time is at least doubled and then check the results again.
- Reconfigure the Phi X 174 simulation into multiple systems containing only 2 molecules of either type to speed up run times and simplify evaluation of the results.
- Reconfigure the Phi X 174 simulation into multiple geometries.
- Perform multiple iterations for all of the above.

With no consideration for any potential peripheral value of this research the basic concepts of verification and validation provide ample justification for these follow-up investigations.

Experiment #2 needs many more iterations performed to fully characterize the periodicity and determine conclusively the source. Since there is no precedent for this type of research these results provide invaluable insight to the theorized concept of intermolecular harmonized resonant vibration and how it pertains to MM simulation. In light of this new insight six specific and pressing questions now need to be answered. They are:

- Does significant pressure variation occur between other types of molecules?
- Is it unique to the NAMD simulator?
- Is it sensitive to sequence or sequence length/repeats?
- Are the pressure variations related to normal modes only?
- What is going on in other slabs of each system?
- What about the basic size of each system? Would a change in system size (amount of solvent) affect the spectrum?
- Why is the DC component so large in only the Parallel and Skew configurations?

In final summary, the flagship hypotheses $\mathbf{H}_{(\text{Resonance NULL})}$ and $\mathbf{H}_{(\text{Harmonized Resonance NULL})}$ stand for now. $\mathbf{H}_{(\text{Simulate Observed Closure NULL})}$ was unexpectedly rejected with caveats. As a direct result $\mathbf{H}_{(\text{Resonance Causes Closure NULL})}$ was not successfully tested because we were unable to simulate closure. $\mathbf{H}_{(\text{Interacting Harmonized Resonance NULL})}$ was refined into $\mathbf{H}_{(\text{Sequence Relationship NULL}),}$ tested and remains inconclusive. Although a direct link between homologous pairing and harmonized inter-molecular vibrations could not be conclusively established with these particular simulations, the theorized existence of harmonized inter-molecular vibrations was not ruled out. The magnitude of the identified periodic behavior found in simulated pressures along with the

inability to conclusively link that behavior to DNA sequence effectively charts a clear course for future research.

From either the biological or modeling perspective, until DNA segregation can be reproduced in simulation and the exact source of the periodicity in simulated pressures is identified molecular mechanical simulation cannot as yet be considered a suitable tool for investigation of Homologous Chromosome Pairing.

APPENDIX-A: ALTERNATE 16 CORE CLUSTER SPECIFICATIONS

1x SUPERMICRO MBD-X8DTL-iF-O Dual LGA 1366 Intel 5500 ATX Dual Intel Xeon 5500 and 5600 Series Server Motherboard

2 x Intel Xeon E5630 Westmere 2.53GHz LGA 1366 80W Quad-Core Server Processor BX80614E5630

2 x Kingston 8GB (2 x 4GB) 240-Pin DDR3 SDRAM DDR3 1333 ECC Registered Server Memory Model KVR1333D3D4R9SK2/8G

2 x Western Digital VelociRaptor WD3000HLFS 300GB 10000 RPM SATA 3.0Gb/s 3.5" Internal Hard Drive -Bare Drive

1 x OCZ StealthXStream OCZ700SXS 700W ATX12V / EPS12V SLI Ready CrossFire Ready Active PFC Power Supply

2 x Intel BXSTS100A Active heat sink with fixed fan

APPENDIX-B: PhiX 174 TABULATED MOVEMENT DATA

Step	N1N2DELTN3 N4	N1N2DELTN5 N6	N1N2DELTN7 N8	N3N4DELTN5 N6	N3N4DELTN7 N8	N5N6DELTN7 N8
1	0.28757	0.316363	0.29762	0.406964	0.299196	0.093495
2	0.166592	0.085073	0.154551	0.076688	0.10685	-0.03151
3	0.222058	0.130968	0.249749	0.147499	0.136212	0.07417
4	-0.0403	0.112229	-0.02059	0.121237	0.050342	0.018387
5	0.061533	-0.00461	0.032386	0.228722	0.165395	0.147079
6	-0.02901	0.065387	0.291235	0.126868	0.054696	0.50009
7	0.19399	0.064281	0.11994	0.056203	0.039473	-0.04864
8	0.055158	0.060403	-0.08437	0.214228	-0.1425	0.212503
9	-0.01792	0.251866	0.127901	0.156992	0.239143	-0.07083
10	-0.07273	-0.25458	-0.053	-0.10431	0.026913	0.080718
11	0.096935	0.251195	0.144863	0.079596	-0.02229	-0.0093
12	-0.05819	-0.01044	-0.08592	0.116929	-0.10162	0.218979
13	0.057616	0.194144	0.058976	0.181762	0.081312	0.003597
14	0.01426	0.204095	0.100002	-0.05545	0.054332	-0.21291
15	0.195778	-0.00194	0.087825	0.250952	0.100816	0.186126
16	0.188237	0.142695	0.204259	-0.08462	-0.18711	0.026835
17	-0.14998	-0.01496	-0.10242	0.164633	0.233521	0.017784
18	-0.08322	0.030187	-0.23051	0.111553	-0.15402	0.044728
19	0.074143	-0.12649	-0.2039	0.077896	-0.19807	0.07868
20	-0.1238	0.143702	0.014621	-0.19022	-0.26335	-0.0008
21	0.176498	0.157551	0.123172	-0.02807	-0.1218	-0.07587
22	-0.00951	-0.08115	-0.08189	0.047757	0.029061	0.014247
23	-0.15319	-0.13895	-0.20347	-0.08452	-0.11436	-0.00089
24	0.243334	-0.11351	-0.05474	0.086701	-0.07384	-0.00776
25	0.020755	0.060144	0.03264	-0.05075	0.07891	-0.18452
26	-0.07389	0.107445	0.033413	-0.11113	-0.21517	0.072424
27	0.130281	0.027222	0.057931	-0.05692	0.107677	-0.26278
28	0.218475	-0.14968	0.029565	0.029919	0.026211	-0.00952
29	0.032664	0.023158	0.088598	-0.11148	-0.05557	-0.03818
30	0.094232	-0.2395	0.011094	-0.15324	-0.02163	-0.03939
31	0.002915	0.076418	0.191674	-0.08874	0.088314	-0.0237
32	0.007479	-0.14642	-0.08684	-0.31435	-0.28987	-0.13357
33	0.193284	-0.05935	0.079793	-0.04012	-0.0662	-0.01327
34	0.057013	-0.07926	0.01427	-0.10905	-0.05054	-0.05694
35	0.125337	0.030691	-0.16331	0.184736	-0.03533	-0.08748
36	-0.08699	-0.12111	-0.0804	-0.20436	-0.12301	-0.0643
37	-0.00674	0.320531	0.045969	0.240787	-0.02814	0.123254
38	0.074271	-0.14916	0.268836	-0.15998	0.15306	0.066641

39	-0.08391	-0.0429	-0.12715	0.186587	0.03279	0.175574
40	-0.01849	-0.01342	0.192797	-0.16686	0.135093	-0.06073
41	0.109417	0.149164	-0.02085	0.075129	-0.1216	-0.05739
42	0.022249	0.023354	0.113665	-0.13372	0.004845	-0.07443
43	-0.07088	-0.23844	-0.14463	-0.01981	0.100043	-0.01893
44	-0.06936	-0.11089	0.049907	-0.18748	-0.0443	0.030224
45	0.291549	-0.13393	0.047689	0.022407	0.017382	-0.07333
46	0.129793	0.005674	0.108413	0.02707	0.144308	-0.07484
47	-0.05544	-0.11395	-0.26077	0.182277	0.072852	-0.0226

APPENDIX-C: PHIX 174 CENTER OF MASS POSITIONAL DATA IN ANGSTROMS

	Position											
Frame	N1N2 X	N1N2 Y	N1N2 Z	N3N4 X	N3N4 Y	N3N4 Z	N5N6 X	N5N6 Y	N5N6 Z	N7N8 X	N7N8 Y	N7N8 Z
0	0.660027	0.670212	495.0429	39.31093	0.596607	495.1824	0.755737	39.47866	495.1485	39.40637	39.34378	495.1896
1(49)	0.429728	0.610081	494.9691	39.36698	0.321619	495.1616	0.769229	39.7331	495.1814	39.51189	39.36783	495.1716
2(99)	0.355006	0.466371	494.9026	39.45967	0.265629	495.0317	0.74949	39.67427	495.0468	39.46148	39.41888	495.1104
3	0.249506	0.444655	495.0764	39.57618	0.215045	494.9849	0.816633	39.78166	495.1154	39.60244	39.50365	495.2549
4	0.323088	0.397258	495.0726	39.60931	0.153805	494.9591	0.80684	39.84731	495.2329	39.61045	39.49342	495.1046
5	0.37641	0.448491	495.0721	39.72378	0.131138	495.0272	0.666734	39.89583	495.2331	39.61838	39.63576	495.2302
6	0.418487	0.306617	495.2478	39.73755	0.156673	495.0882	0.398203	39.82056	495.3657	39.85063	39.71586	495.3109
7	0.442044	0.217854	495.2436	39.95545	0.177191	495.1078	0.491483	39.79522	495.5252	39.89462	39.77626	495.2707
8	0.429116	0.187812	495.2228	39.99787	0.174331	495.1522	0.263502	39.82301	495.7313	39.87791	39.63046	495.3765
9	0.333105	0.130953	495.2233	39.8838	0.090444	495.1023	0.20771	40.01759	495.7766	39.75068	39.78534	495.3795
10	0.367011	0.356494	495.2302	39.84357	0.02759	495.0916	0.350732	39.98658	495.9198	39.97084	39.75005	495.2568
11	0.362715	0.140628	495.0979	39.9377	0.063245	495.041	0.330061	40.01797	495.9881	39.94243	39.762	495.4346
12	0.411296	0.138145	495.0181	39.92764	-0.02681	495.1527	0.226949	40.00315	495.9725	40.05694	39.5711	495.4293
13	0.379754	0.048137	495.097	39.95414	-0.05569	495.1203	0.131629	40.10876	495.9744	39.96725	39.62064	495.6893
14	0.335765	-0.05227	495.147	39.92418	-0.08265	494.9754	0.317218	40.21845	495.7399	39.93967	39.64666	495.6317
15	0.262868	-0.05453	495.0827	40.04699	-0.14384	494.916	0.141084	40.20971	495.9209	39.94941	39.6852	495.6298
16	0.174034	-0.11593	495.2046	40.14598	0.009722	494.9748	0.288715	40.29225	495.9832	40.12219	39.65165	495.6928
17	0.329839	-0.0391	495.3438	40.14956	-0.14884	494.8588	0.286787	40.35692	495.9691	40.13611	39.72851	495.465
18	0.4385	-0.01365	495.3988	40.17435	-0.13623	494.8661	0.194071	40.41457	495.8117	40.08669	39.58601	495.5329
19	0.3781	0.136642	495.4444	40.18727	-0.10624	494.8913	0.087471	40.43953	495.6724	40.05502	39.41841	495.5204
20	0.493521	-0.019	495.5832	40.17793	-0.06201	494.9216	0.293545	40.42852	495.731	40.25448	39.19859	495.5988
21	0.474606	-0.21887	495.4861	40.3333	0.006257	494.7328	0.38017	40.38637	495.685	40.26414	39.145	495.4136
22	0.594468	-0.14774	495.5789	40.44188	-0.00794	494.7191	0.428104	40.37597	495.8049	40.32698	39.15735	495.5267
23	0.716015	-0.03023	495.5896	40.40994	0.055219	494.7106	0.431052	40.35379	495.8289	40.32718	39.10792	495.431
24	0.608001	0.130225	495.6399	40.54514	0.069439	494.7508	0.472774	40.4022	495.6799	40.35901	39.04804	495.4649
25	0.619932	-0.03522	495.6057	40.57278	0.046278	494.5143	0.493418	40.29694	495.5865	40.19917	39.10234	495.2348
26	0.590082	-0.02053	495.5565	40.47195	0.119391	494.5846	0.598108	40.41927	495.5864	40.36731	38.96369	495.2008
27	0.602645	-0.13243	495.414	40.61375	0.046777	494.405	0.81069	40.33396	495.5067	40.32169	38.99467	495.1952
28	0.49887	0.047985	495.3925	40.72857	0.103377	494.3711	0.853872	40.36378	495.3974	40.35796	39.07501	495.2458
29	0.585555	-0.13774	495.1321	40.85204	0.171078	494.3443	0.903569	40.20079	495.3757	40.37445	39.08743	495.1509
30	0.44737	0.090383	495.0011	40.81104	0.137516	494.3104	1.096681	40.18505	495.2977	40.52737	39.03375	495.1352
31	0.429122	-0.09212	494.8342	40.79684	0.050111	494.2285	1.187242	40.07627	495.2242	40.59726	39.03511	495.0576
32	0.337969	0.068626	494.8398	40.71082	0.280301	494.1103	1.327288	40.08598	495.1786	40.60196	38.97218	495.0913
33	0.257108	0.043332	494.7267	40.82421	0.285146	494.061	1.404177	39.9954	495.2276	40.6659	38.91142	495.0103
34	0.137914	0.110829	494.7043	40.76173	0.30345	494.0047	1.45917	39.9798	495.0563	40.66401	38.87668	495.0578
35	0.226915	0.065627	494.5679	40.97684	0.183449	493.8966	1.52065	39.96486	495.0499	40.6336	38.71969	494.9583
36	0.212009	0.144848	494.6364	40.87063	0.188868	493.7348	1.655843	39.9193	494.9642	40.70182	38.59472	495.0688
37	0.316114	-0.0893	494.662	40.96717	0.101019	493.7425	1.591795	40.01188	494.9545	40.7538	38.48357	494.9212

38	0.228704	-0.08495	494.6119	40.95665	0.098773	493.8142	1.656552	39.86155	494.9427	40.89589	38.63613	494.9554
39	0.400719	-0.00994	494.6419	41.04182	0.113015	493.6974	1.50873	39.90398	494.9487	40.92368	38.67804	494.9968
40	0.224425	0.101912	494.5712	40.8496	0.048205	493.7373	1.701313	39.99124	494.7477	41.05336	38.74588	495.0996
41	0.294497	-0.08716	494.6256	41.02908	0.03839	493.7979	1.742941	39.95266	494.7555	41.0357	38.62106	494.9738
42	0.258656	-0.12817	494.7442	41.01539	0.081213	493.9284	1.851684	39.92964	494.8291	41.07205	38.6691	495.0909
43	0.369563	0.075637	494.7949	41.0504	0.154508	493.7414	1.816524	39.90043	494.8613	41.02255	38.8305	495.2516
44	0.179671	0.232041	494.6721	40.79403	0.127447	493.7397	1.889089	39.9353	494.7983	41.1198	38.75258	495.3766
45	0.134802	0.216242	494.677	41.04048	0.136324	493.7287	1.936496	39.78151	494.7353	41.09882	38.78203	495.3215
46	0.120695	0.191152	494.8262	41.15559	0.102707	493.8527	2.019192	39.75745	494.7102	41.10671	38.89156	495.4768
47	0.281238	0.327646	494.7843	41.26004	0.060803	493.8164	1.943123	39.7907	494.7928	41.0095	38.92036	495.4745

APPENDIX-D: NAMD SIMULATION PARAMETER FILES (TEMPLATES)

Minimize.conf

(This is the minimization config file for the linear system. It served as a template for all minimizations.)

```
#############################################################
## JOB DESCRIPTION                                         ##
#############################################################
# Minimization step 1
# tataa_80_ionized with 2 linear molecules matching sequence ensembles
#############################################################
## ADJUSTABLE PARAMETERS                                   ##
#############################################################
structure          /home/rick/NAMD/20Mer/Bigger_box/tataaa_80_ionized.psf
coordinates        /home/rick/NAMD/20Mer/Bigger_box/tataaa_80_ionized.pdb
set temperature      0
set outputname      /home/rick/NAMD/20Mer/Bigger_box/runs/tataaa_80_min
firsttimestep      0
#############################################################
## SIMULATION PARAMETERS                                   ##
#############################################################
# IMD settings for VMD interface
if {1}  {
IMDon          on
IMDport 3000
IMDfreq 1
IMDwait no
            }
# Input
paraTypeCharmm          on
parameters              /home/rick/NAMD/20Mer/par_all27_na.prm
temperature            $temperature
# Force-Field Parameters
exclude             scaled1-4
1-4scaling          1.0
cutoff              12.
switching           on
switchdist          10.
pairlistdist        13.5
# Integrator Parameters
timestep            1.0  # 1fs/step
nonbondedFreq       1
fullElectFrequency  2
stepspercycle       10
# Periodic Boundary Conditions
#>Main< (20Mer_AT_GC) 63 % measur center $everyone
#-0.005743197165429592 0.07249268144369125 72.2677993774414
#>Main< (20Mer_AT_GC) 64 % measure minmax $everyone
#{-16.802000045776367 -17.200000762939453 -10.27400016784668} {16.798999786376953
17.270999908447266 154.52200317382813}
cellBasisVector1     33.601   0.0    0.0
cellBasisVector2     0.0    34.47   0.0
cellBasisVector3     0.0     0.0    164.8
cellOrigin          -0.006   0.072  72.268
wrapAll             on
wrapNearest              yes
COMmotion                no
# PME (for full-system periodic electrostatics)
PME                 yes
PMEGridSpacing     1
# Output
#restartfreq         400  # 1000steps = every 1ps
outputName           $outputname
```

```
dcdfreq             100
xstFreq             100
outputEnergies      100
outputPressure      100
#outputTiming       20
#############################################################
## EXECUTION SCRIPT                                        ##
#############################################################
# Minimization
minimize          10000
```

Equilibrate.conf

(This is the equilibration config file for the linear system. It served as a template for all minimizations.)

```
#############################################################
## JOB DESCRIPTION                                         ##
#############################################################
# Equilibration Step 2 allow heated system to relax 2000 steps at 310 deg
# 2mol_linear_ionized with 2 linear molecules matching sequence ensembles
#############################################################
## ADJUSTABLE PARAMETERS                                   ##
#############################################################
structure          /home/rick/NAMD/20Mer/Bigger_box/tataaa_80_ionized.psf
coordinates        /home/rick/NAMD/20Mer/Bigger_box/tataaa_80_ionized.pdb
bincoordinates /home/rick/NAMD/20Mer/Bigger_box/runs/tataaa_80_heat.coor
extendedSystem /home/rick/NAMD/20Mer/Bigger_box/runs/tataaa_80_heat.xsc
binvelocities  /home/rick/NAMD/20Mer/Bigger_box/runs/tataaa_80_heat.vel
set outputname /home/rick/NAMD/20Mer/Bigger_box/runs/tataaa_80_equilibrated
set temperature       310
#############################################################
## SIMULATION PARAMETERS                                   ##
#############################################################
#Margin setting
margin                 2.5
# Input
paraTypeCharmm          on
parameters              /home/rick/NAMD/20Mer/par_all27_na.prm
# Constant Pressure Control (variable volume)
if {1} {
useGroupPressure      yes ;# needed for 2fs steps
useFlexibleCell       no  ;# no for water box, yes for membrane
useConstantArea       no  ;# no for water box, yes for membrane
langevinPiston        on
langevinPistonTarget  1.01325 ;#  in bar -> 1 atm
langevinPistonPeriod  100.
langevinPistonDecay   50.
langevinPistonTemp    $temperature
}
# Constant Temperature Control
langevin            on    ;# do langevin dynamics
langevinDamping     5     ;# damping coefficient (gamma) of 5/ps
langevinTemp        $temperature
langevinHydrogen    no    ;# don't couple langevin bath to hydrogens
# Force-Field Parameters
exclude             scaled1-4
1-4scaling          1.0
cutoff              12.
switching           on
switchdist          10.
pairlistdist        13.5
# Integrator Parameters
```

```
timestep            2.0  # 1fs/step
nonbondedFreq       1
fullElectFrequency  2
stepspercycle       10
wrapAll             on
wrapNearest             yes
COMmotion               no
# PME (for full-system periodic electrostatics)
PME                 yes
PMEGridSpacing     1
# Output
outputName            $outputname
dcdfreq             100
xstFreq             100
outputEnergies      100
outputPressure      100
outputTiming     100
#############################################################
## EXECUTION SCRIPT                                        ##
#############################################################
# Basic equilibration
numsteps       10000
```

Heat.conf

(This is the configuration file for a typical MD heating process of the Linear configuration. This file served as a template for the remaining 4 runs of the linear configuration and all 5 runs for all other simulations.)

```
#############################################################
## JOB DESCRIPTION                                         ##
#############################################################
# Equilibration Step 1 slowly heat to 310 deg K or 98 deg F
# 2mol_linear_ionized with 2 linear molecules matching sequence ensembles


#############################################################
## ADJUSTABLE PARAMETERS                                   ##
#############################################################
structure          /home/rick/NAMD/20Mer/Bigger_box/tataaa_80_ionized.psf
coordinates        /home/rick/NAMD/20Mer/Bigger_box/tataaa_80_ionized.pdb
bincoordinates /home/rick/NAMD/20Mer/Bigger_box/runs/tataaa_80_min.coor
extendedSystem /home/rick/NAMD/20Mer/Bigger_box/runs/tataaa_80_min.xsc
set outputname /home/rick/NAMD/20Mer/Bigger_box/runs/tataaa_80_heat
#############################################################
## SIMULATION PARAMETERS                                   ##
#############################################################
# Input
paraTypeCharmm          on
parameters              /home/rick/NAMD/20Mer/par_all27_na.prm
temperature            0
reassignFreq   1
reassignTemp   0
reassignIncr   1
reassignHold   310
# Force-Field Parameters
exclude             scaled1-4
1-4scaling          1.0
cutoff              12.
```

```
switching           on
switchdist          10.
pairlistdist        13.5
# Integrator Parameters
timestep            1.0  # 1fs/step
nonbondedFreq       1
fullElectFrequency  2
stepspercycle       10
wrapAll             on
wrapNearest              yes
COMmotion                no
# PME (for full-system periodic electrostatics)
PME                 yes
PMEGridSpacing     1
# Output
outputName            $outputname
dcdfreq             100
xstFreq             100
outputEnergies      100
outputPressure      100
outputTiming     100
#############################################################
## EXECUTION SCRIPT                                        ##
#############################################################
# Incremental Heating
numsteps       500
```

Linear_21slab_run1.conf

(This is the configuration file for a typical production MD run of the Linear configuration. This file served as a template for the remaining 4 runs of the linear configuration and all 5 runs for all other simulations.)

```
#############################################################
## JOB DESCRIPTION                                         ##
#############################################################
# Production Run 1 run 200000 steps with profile pressures output no pressure control no temp
control
#no rigid bonds is defaulted but using 21 pp slabs for higher resolution in matlab
# 2mol_linear_ionized with 2 linear molecules matching sequence ensembles
#############################################################
## ADJUSTABLE PARAMETERS                                   ##
#############################################################
structure          /home/rick/NAMD/Linear/tataaa_80_ionized.psf
coordinates        /home/rick/NAMD/Linear/tataaa_80_ionized.pdb
bincoordinates /home/rick/NAMD/Linear/runs/tataaa_80_equilibrated.coor
extendedSystem /home/rick/NAMD/Linear/runs/tataaa_80_equilibrated.xsc
binvelocities  /home/rick/NAMD/Linear/runs/tataaa_80_equilibrated.vel
firsttimestep  0
set outputname /home/rick/NAMD/Linear/runs/prod/run1/Linear_21slab_run1
set temperature        310
#############################################################
## SIMULATION PARAMETERS                                   ##
#############################################################
# IMD settings for VMD interface
if {1}  {
IMDon          on
IMDport 3001
```

```
IMDfreq 1
IMDwait no
             }
# Input
paraTypeCharmm          on
parameters              /home/rick/NAMD/20Mer/par_all27_na.prm
# Constant Temperature Control no
if {1}  {
langevin            on    ;# do langevin dynamics
langevinDamping     5     ;# damping coefficient (gamma) of 5/ps
langevinTemp        $temperature
langevinHydrogen    no    ;# don't couple langevin bath to hydrogens
}
# Constant Pressure Control (variable volume) no pressure influence wanted
if {1} {
#useGroupPressure      yes ;# needed for 2fs steps
useFlexibleCell       no  ;# no for water box, yes for membrane
useConstantArea       no  ;# no for water box, yes for membrane
langevinPiston        on
langevinPistonTarget  1.01325 ;#  in bar -> 1 atm
langevinPistonPeriod  200.
langevinPistonDecay   100.
langevinPistonTemp    $temperature
}
useGroupPressure      no ;# needed for 2fs steps
# Force-Field Parameters
exclude             scaled1-4
1-4scaling          1.0
cutoff              12.
switching           on
switchdist          10.
pairlistdist        13.5
# Integrator Parameters
timestep            2.0  # 1fs/step
rigidBonds          none  # all needed for 2fs steps
nonbondedFreq       1
fullElectFrequency  2
stepspercycle       10
seed                   05241986
wrapAll             on
wrapNearest             yes
# PME (for full-system periodic electrostatics)
PME                 yes
PMEGridSpacing     1
#Pressure Profile Output
if {1}  {
pressureProfile               on
pressureProfileSlabs   21
pressureProfileFreq    10
             }
# Output
outputName            $outputname
dcdfreq             10
xstFreq             1000
outputEnergies      1000
outputPressure      1000
outputTiming       100
##############################################################
## EXECUTION SCRIPT                                         ##
##############################################################
# Production Run with pressure profile output
numsteps       199999
```

Get_Pressures_run1.conf

(This is the configuration file for a typical production MD RE-run of the Linear configuration where molecular dynamics are not run. The original DCD trajectory file is used in this configuration to calculate offline pressures only. This file served as a template for the remaining 4 runs of the linear configuration and all 5 runs for all other simulations.)

```
#############################################################
## Get Ewald Pressures                                     ##
#############################################################
# Production Run 1 run 200000 steps with profile pressures output no pressure control no temp
control
#no rigid bonds is defaulted but using 21 pp slabs for higher resolution in matlab
# 2mol_linear_ionized with 2 linear molecules matching sequence ensembles
#############################################################
## ADJUSTABLE PARAMETERS                                   ##
#############################################################
structure          /home/rick/NAMD/Linear/tataaa_80_ionized.psf
coordinates        /home/rick/NAMD/Linear/tataaa_80_ionized.pdb
bincoordinates /home/rick/NAMD/Linear/runs/tataaa_80_equilibrated.coor
extendedSystem /home/rick/NAMD/Linear/runs/tataaa_80_equilibrated.xsc
binvelocities  /home/rick/NAMD/Linear/runs/tataaa_80_equilibrated.vel
#firsttimestep 0
set outputname /home/rick/NAMD/Linear/runs/prod/run1/tataaa_80_200000_run1_Ewald_Pressure
set temperature        310
#############################################################
## SIMULATION PARAMETERS                                   ##
#############################################################
# IMD settings for VMD interface
if {1}  {
IMDon          on
IMDport 3001
IMDfreq 1
IMDwait no
               }
# Input
paraTypeCharmm          on
parameters              /home/rick/NAMD/20Mer/par_all27_na.prm
# Constant Temperature Control no
if {1}  {
langevin            on    ;# do langevin dynamics
langevinDamping     5     ;# damping coefficient (gamma) of 5/ps
langevinTemp        $temperature
langevinHydrogen    no    ;# don't couple langevin bath to hydrogens
}
# Constant Pressure Control (variable volume) no pressure influence wanted
if {1} {
#useGroupPressure      yes ;# needed for 2fs steps
useFlexibleCell       no  ;# no for water box, yes for membrane
useConstantArea       no  ;# no for water box, yes for membrane
langevinPiston        on
langevinPistonTarget  1.01325 ;#  in bar -> 1 atm
langevinPistonPeriod  100.
langevinPistonDecay   50.
langevinPistonTemp    $temperature
}
useGroupPressure      no ;# needed for 2fs steps
# Force-Field Parameters
exclude             scaled1-4
1-4scaling          1.0
cutoff              12.
switching           on
```

```
switchdist          10.
pairlistdist        13.5
# Integrator Parameters
timestep            2.0  # 1fs/step
rigidBonds          none  # all needed for 2fs steps
nonbondedFreq       1
fullElectFrequency  2
stepspercycle       10
seed                   05241986
wrapAll             on
wrapNearest             yes
# PME (for full-system periodic electrostatics)
PME                 yes
PMEGridSpacing     1
# Output
outputName            $outputname
#dcdfreq             1000
#xstFreq             1000
#outputEnergies      1000
#outputPressure      1000
#outputTiming    1000
#Pressure Profile Output
if {1}  {
pressureProfile               on
pressureProfileSlabs   21
pressureProfileFreq    10
pressureProfileEwald   on
pressureProfileEwaldX  10
pressureProfileEwaldY  10
pressureProfileEwaldZ  10
               }
set ts 0
firstTimestep $ts
coorfile open dcd /home/rick/NAMD/Linear/runs/prod/run1/Linear_21slab_run1.dcd
while { [coorfile read] != -1} {
incr ts 10
firstTimestep $ts
run 0
                     }
#############################################################
## EXECUTION SCRIPT                                        ##
#############################################################
# Production Run with pressure profile output
```

APPENDIX-E: PERL FILES FOR PARSING DATA

Parse_Ewald_Pressures_32_33_34.pl (linear)

Parse_Ewald_Pressures_11_12_13.pl (parallel)

Parse_Ewald_Pressures_14_15_16.pl (perpt)

Parse_Ewald_Pressures_11_12_13.pl (skew)

Parse_Ewald_Pressures_32_33_34.pl (random)

(This is a series of programs used to open the NAMD log file outputs from each of the geometric configuration files for experiment #2. They would find and load the pressure profile data specifically for the appropriate slab and write it out to a summary file for later summation. The exact code from Parse_Ewald_Pressures_32_33_34.pl for the original linear system is listed below.)

```
#############################################################3
##
## 21 slab cell midway is slab 11
## Corresponding to string position 32 33 34 with time at position 1
##
## 7 slab cell midway is where?
## Coressponding to string position 11 12 13 with time at position 1
##
##  Slab# times 3 minus 1 ie 4*3=12 12-1=11  positions are 11 12 13
##
$filelocation = "N:\\Linear\\runs\\prod\\run1\\tataaa_80_run1_ewald_pressure.log";
$xdata = ">N:\\Linear\\runs\\prod\\run1\\x_ewald_pressures.txt";
$ydata = ">N:\\Linear\\runs\\prod\\run1\\y_ewald_pressures.txt";
$zdata = ">N:\\Linear\\runs\\prod\\run1\\z_ewald_pressures.txt";
$combined_xdata = ">N:\\Linear\\runs\\prod\\run_combined\\x_ewald_pressures.txt";
$combined_ydata = ">N:\\Linear\\runs\\prod\\run_combined\\y_ewald_pressures.txt";
$combined_zdata = ">N:\\Linear\\runs\\prod\\run_combined\\z_ewald_pressures.txt";
open (INFILE, $filelocation);
open (OUTX, $xdata);
open (OUTY, $ydata);
open (OUTZ, $zdata);
open (OUTCOMBINEDX, $combined_xdata);
open (OUTCOMBINEDY, $combined_ydata);
open (OUTCOMBINEDZ, $combined_zdata);
while (<INFILE>)                {
@values = split(/ /,$_);
#print stdout ($values[0]);
if ($values[0] eq "PRESSUREPROFILE\:") {
#$time = ($values[1]* .0000000000000001);
$time = $values[1];
$stringx = $time.",".$values[32];
$stringy = $time.",".$values[33];
$stringz = $time.",".$values[34];
print OUTX $stringx . "\n";
print OUTY $stringy . "\n";
print OUTZ $stringz . "\n";
print OUTCOMBINEDX $stringx . "\n";
print OUTCOMBINEDY $stringy . "\n";
print OUTCOMBINEDZ $stringz . "\n";

         }
                        }
```

```
close INFILE;
close OUTX;
close OUTY;
close OUTZ;

print "all done Run1"."\n";
#################################################################3
##
##
##
$filelocation = "N:\\Linear\\runs\\prod\\run2\\Get_ewald_pressures_run2.log";
$xdata = ">N:\\Linear\\runs\\prod\\run2\\x_ewald_pressures.txt";
$ydata = ">N:\\Linear\\runs\\prod\\run2\\y_ewald_pressures.txt";
$zdata = ">N:\\Linear\\runs\\prod\\run2\\z_ewald_pressures.txt";
open (INFILE, $filelocation);
open (OUTX, $xdata);
open (OUTY, $ydata);
open (OUTZ, $zdata);
while (<INFILE>)                {
@values = split(/ /,$_);
#print stdout ($values[0]);
if ($values[0] eq "PRESSUREPROFILE\:") {
#$time = ($values[1]* .0000000000000001);
$time = $values[1];
$stringx = $time.",".$values[32];
$stringy = $time.",".$values[33];
$stringz = $time.",".$values[34];
print OUTX $stringx . "\n";
print OUTY $stringy . "\n";
print OUTZ $stringz . "\n";
print OUTCOMBINEDX $stringx . "\n";
print OUTCOMBINEDY $stringy . "\n";
print OUTCOMBINEDZ $stringz . "\n";

          }

                                }
close INFILE;
close OUTX;
close OUTY;
close OUTZ;

print  "all done Run2"."\n";
#################################################################3
##
##
##
$filelocation = "N:\\Linear\\runs\\prod\\run3\\Get_ewald_pressures_run3.log";
$xdata = ">N:\\Linear\\runs\\prod\\run3\\x_ewald_pressures.txt";
$ydata = ">N:\\Linear\\runs\\prod\\run3\\y_ewald_pressures.txt";
$zdata = ">N:\\Linear\\runs\\prod\\run3\\z_ewald_pressures.txt";
open (INFILE, $filelocation);
open (OUTX, $xdata);
open (OUTY, $ydata);
open (OUTZ, $zdata);
while (<INFILE>)                {
@values = split(/ /,$_);
#print stdout ($values[0]);
if ($values[0] eq "PRESSUREPROFILE\:") {
#$time = ($values[1]* .0000000000000001);
$time = $values[1];
$stringx = $time.",".$values[32];
$stringy = $time.",".$values[33];
$stringz = $time.",".$values[34];
print OUTX $stringx . "\n";
print OUTY $stringy . "\n";
print OUTZ $stringz . "\n";
print OUTCOMBINEDX $stringx . "\n";
print OUTCOMBINEDY $stringy . "\n";
print OUTCOMBINEDZ $stringz . "\n";

          }

                                }
close INFILE;
close OUTX;
close OUTY;
```

```
close OUTZ;

print  "all done Run3"."\n";
#############################################################3
##
##
##
$filelocation = "N:\\Linear\\runs\\prod\\run4\\Get_ewald_pressures_run4.log";
$xdata = ">N:\\Linear\\runs\\prod\\run4\\x_ewald_pressures.txt";
$ydata = ">N:\\Linear\\runs\\prod\\run4\\y_ewald_pressures.txt";
$zdata = ">N:\\Linear\\runs\\prod\\run4\\z_ewald_pressures.txt";
open (INFILE, $filelocation);
open (OUTX, $xdata);
open (OUTY, $ydata);
open (OUTZ, $zdata);
while (<INFILE>)                {
@values = split(/ /,$_);
#print stdout ($values[0]);
if ($values[0] eq "PRESSUREPROFILE\:") {
#$time = ($values[1]* .0000000000000001);
$time = $values[1];
$stringx = $time.",".$values[32];
$stringy = $time.",".$values[33];
$stringz = $time.",".$values[34];
print OUTX $stringx . "\n";
print OUTY $stringy . "\n";
print OUTZ $stringz . "\n";
print OUTCOMBINEDX $stringx . "\n";
print OUTCOMBINEDY $stringy . "\n";
print OUTCOMBINEDZ $stringz . "\n";

          }

                                  }
close INFILE;
close OUTX;
close OUTY;
close OUTZ;

print  "all done Run4"."\n";
#############################################################3
##
##
##
$filelocation = "N:\\Linear\\runs\\prod\\run5\\Get_ewald_pressures_run5.log";
$xdata = ">N:\\Linear\\runs\\prod\\run5\\x_ewald_pressures.txt";
$ydata = ">N:\\Linear\\runs\\prod\\run5\\y_ewald_pressures.txt";
$zdata = ">N:\\Linear\\runs\\prod\\run5\\z_ewald_pressures.txt";
open (INFILE, $filelocation);
open (OUTX, $xdata);
open (OUTY, $ydata);
open (OUTZ, $zdata);
while (<INFILE>)                {
@values = split(/ /,$_);
#print stdout ($values[0]);
if ($values[0] eq "PRESSUREPROFILE\:") {
#$time = ($values[1]* .0000000000000001);
$time = $values[1];
$stringx = $time.",".$values[32];
$stringy = $time.",".$values[33];
$stringz = $time.",".$values[34];
print OUTX $stringx . "\n";
print OUTY $stringy . "\n";
print OUTZ $stringz . "\n";
print OUTCOMBINEDX $stringx . "\n";
print OUTCOMBINEDY $stringy . "\n";
print OUTCOMBINEDZ $stringz . "\n";

          }

                                  }
close INFILE;
close OUTX;
close OUTY;
close OUTZ;
close OUTCOMBINEDX;
close OUTCOMBINEDY;
```

```
close OUTCOMBINEDZ;

print  "all done Run5 and Combined";
```

Parse_Runtime_Pressures_32_33_34.pl (linear)

Parse_Runtime_Pressures_11_12_13.pl (parallel)

Parse_Runtime_Pressures_14_15_16.pl (perpt)

Parse_Runtime_Pressures_11_12_13.pl (skew)

Parse_Runtime_Pressures_32_33_34.pl (random)

(This also is a series of programs used to open the NAMD log file outputs from each of the geometric configuration files for experiment #2. They are virtually identical to the Ewald files except they would find and load the pressure profile data from the runtime logs instead of the Ewald calculation logs and write it out to a summary file for later summation. The exact code from Parse_Runtime_Pressures_32_33_34.pl for the original linear system is listed below.)

```
#################################################################3
##
## 21 slab cell midway is slab 11
## Corresponding to string position 32 33 34 with time at position 1
## Skip first pressure every run

#$filelocation = "N:\\Linear\\runs\\prod\\run1\\Linear_21slab_run1.log";
#$xdata = ">N:\\Linear\\runs\\prod\\run1\\x_runtime_pressures.txt";
#$ydata = ">N:\\Linear\\runs\\prod\\run1\\y_runtime_pressures.txt";
#$zdata = ">N:\\Linear\\runs\\prod\\run1\\z_runtime_pressures.txt";
#$combined_xdata = ">N:\\Linear\\runs\\prod\\run_combined\\x_runtime_pressures.txt";
#$combined_ydata = ">N:\\Linear\\runs\\prod\\run_combined\\y_runtime_pressures.txt";
#$combined_zdata = ">N:\\Linear\\runs\\prod\\run_combined\\z_runtime_pressures.txt";

$filelocation = "N:\\Linear\\runs\\prod\\run1\\Linear_21slab_run1.log";
$xdata = ">N:\\Linear\\runs\\prod\\run1\\x_runtime_pressures.txt";
$ydata = ">N:\\Linear\\runs\\prod\\run1\\y_runtime_pressures.txt";
$zdata = ">N:\\Linear\\runs\\prod\\run1\\z_runtime_pressures.txt";
$combined_xdata = ">N:\\Linear\\runs\\prod\\run_combined\\x_runtime_pressures.txt";
$combined_ydata = ">N:\\Linear\\runs\\prod\\run_combined\\y_runtime_pressures.txt";
$combined_zdata = ">N:\\Linear\\runs\\prod\\run_combined\\z_runtime_pressures.txt";

##########  Establish Counting Routines Here   ##################

open (INFILE, $filelocation);
while (<INFILE>)                  {
@values = split(/ /,$_);
if ($values[0] eq "PRESSUREPROFILE\:") {
$run1_count++;
                                       }
                                   }
close INFILE;
print stdout "Counted Run1:  ".$run1_count."\n";
#################################################################3
##
##
##
$filelocation = "N:\\Linear\\runs\\prod\\run2\\Linear_21slab_run2.log";
```

```
open (INFILE, $filelocation);
while (<INFILE>)                  {
@values = split(/ /,$_);
if ($values[0] eq "PRESSUREPROFILE\:") {
$run2_count++;
                                  }

                                  }
close INFILE;
print stdout "Counted Run2:  ".$run2_count."\n";
#################################################################3
##
##
##
$filelocation = "N:\\Linear\\runs\\prod\\run3\\Linear_21slab_run3.log";
open (INFILE, $filelocation);
while (<INFILE>)                  {
@values = split(/ /,$_);
if ($values[0] eq "PRESSUREPROFILE\:") {
$run3_count++;
                                       }

                                  }
close INFILE;
print stdout "Counted Run3:  ".$run3_count."\n";
#################################################################3
##
##
##
$filelocation = "N:\\Linear\\runs\\prod\\run4\\Linear_21slab_run4.log";
open (INFILE, $filelocation);
while (<INFILE>)                  {
@values = split(/ /,$_);
if ($values[0] eq "PRESSUREPROFILE\:") {
$run4_count++;
                                       }
                                  }
close INFILE;
print stdout "Counted Run4:  ".$run4_count."\n";
#################################################################3
##
##
##
$filelocation = "N:\\Linear\\runs\\prod\\run5\\Linear_21slab_run5.log";
open (INFILE, $filelocation);
while (<INFILE>)                  {
@values = split(/ /,$_);
if ($values[0] eq "PRESSUREPROFILE\:") {
$run5_count++;
                                       }

                                  }
close INFILE;
print stdout "Counted Run5:  ".$run5_count."\n";

##########    End Counting Routines ##############################

$filelocation = "N:\\Linear\\runs\\prod\\run1\\Linear_21slab_run1.log";
$xdata = ">N:\\Linear\\runs\\prod\\run1\\x_runtime_pressures.txt";
$ydata = ">N:\\Linear\\runs\\prod\\run1\\y_runtime_pressures.txt";
$zdata = ">N:\\Linear\\runs\\prod\\run1\\z_runtime_pressures.txt";

open (INFILE, $filelocation);
open (OUTX, $xdata);
open (OUTY, $ydata);
open (OUTZ, $zdata);
open (OUTCOMBINEDX, $combined_xdata);
open (OUTCOMBINEDY, $combined_ydata);
open (OUTCOMBINEDZ, $combined_zdata);
$skip_first = 0;
while (<INFILE>)                  {

@values = split(/ /,$_);
#print stdout ($values[0]);
if ($values[0] eq "PRESSUREPROFILE\:") {
#$time = ($values[1]* .0000000000000001);
$run1_current++;
```

```
$time = $values[1];
$stringx = $time.",".$values[32];
$stringy = $time.",".$values[33];
$stringz = $time.",".$values[34];
if ($run1_current >= $run1_count) {
goto skip_print;
                        }
print OUTX $stringx . "\n";
print OUTY $stringy . "\n";
print OUTZ $stringz . "\n";
print OUTCOMBINEDX $stringx . "\n";
print OUTCOMBINEDY $stringy . "\n";
print OUTCOMBINEDZ $stringz . "\n";
skip_print:

          }
                                  }
close INFILE;
close OUTX;
close OUTY;
close OUTZ;

print stdout "all done Run1"."\n";
#############################################################3
##
##
##
$filelocation = "N:\\Linear\\runs\\prod\\run2\\Linear_21slab_run2.log";
$xdata = ">N:\\Linear\\runs\\prod\\run2\\x_runtime_pressures.txt";
$ydata = ">N:\\Linear\\runs\\prod\\run2\\y_runtime_pressures.txt";
$zdata = ">N:\\Linear\\runs\\prod\\run2\\z_runtime_pressures.txt";
open (INFILE, $filelocation);
open (OUTX, $xdata);
open (OUTY, $ydata);
open (OUTZ, $zdata);
$skip_first = 0;
while (<INFILE>)                  {
@values = split(/ /,$_);
#print stdout ($values[0]);
if ($values[0] eq "PRESSUREPROFILE\:") {
#$time = ($values[1]* .0000000000000001);
$run2_current++;
$time = $values[1];
$stringx = $time.",".$values[32];
$stringy = $time.",".$values[33];
$stringz = $time.",".$values[34];
if ($run2_current >= $run2_count) {
print stdout "Skipped run2 at:  ".$run2_current."\n";
goto skip_2;
                        }
print OUTX $stringx . "\n";
print OUTY $stringy . "\n";
print OUTZ $stringz . "\n";
print OUTCOMBINEDX $stringx . "\n";
print OUTCOMBINEDY $stringy . "\n";
print OUTCOMBINEDZ $stringz . "\n";
skip_2:

          }

                                  }
close INFILE;
close OUTX;
close OUTY;
close OUTZ;

print stdout "all done Run2"."\n";
#############################################################3
##
##
##
$filelocation = "N:\\Linear\\runs\\prod\\run3\\Linear_21slab_run3.log";
$xdata = ">N:\\Linear\\runs\\prod\\run3\\x_runtime_pressures.txt";
$ydata = ">N:\\Linear\\runs\\prod\\run3\\y_runtime_pressures.txt";
$zdata = ">N:\\Linear\\runs\\prod\\run3\\z_runtime_pressures.txt";
open (INFILE, $filelocation);
open (OUTX, $xdata);
```

```
open (OUTY, $ydata);
open (OUTZ, $zdata);
$skip_first = 0;
while (<INFILE>)                {
@values = split(/ /,$_);
#print stdout ($values[0]);
if ($values[0] eq "PRESSUREPROFILE\:") {
$run3_current++;
#$time = ($values[1]* .0000000000000001);
$time = $values[1];
$stringx = $time.",".$values[32];
$stringy = $time.",".$values[33];
$stringz = $time.",".$values[34];
if ($run3_current >= $run3_count) {
goto skip_3;
                        }
print OUTX $stringx . "\n";
print OUTY $stringy . "\n";
print OUTZ $stringz . "\n";
print OUTCOMBINEDX $stringx . "\n";
print OUTCOMBINEDY $stringy . "\n";
print OUTCOMBINEDZ $stringz . "\n";
skip_3:
         }

                                  }
close INFILE;
close OUTX;
close OUTY;
close OUTZ;

print stdout "all done Run3"."\n";
#################################################################3
##
##
##
$filelocation = "N:\\Linear\\runs\\prod\\run4\\Linear_21slab_run4.log";
$xdata = ">N:\\Linear\\runs\\prod\\run4\\x_runtime_pressures.txt";
$ydata = ">N:\\Linear\\runs\\prod\\run4\\y_runtime_pressures.txt";
$zdata = ">N:\\Linear\\runs\\prod\\run4\\z_runtime_pressures.txt";
open (INFILE, $filelocation);
open (OUTX, $xdata);
open (OUTY, $ydata);
open (OUTZ, $zdata);
$skip_first = 0;
while (<INFILE>)                {
@values = split(/ /,$_);
#print stdout ($values[0]);
if ($values[0] eq "PRESSUREPROFILE\:") {
$run4_current++;
#$time = ($values[1]* .0000000000000001);
$time = $values[1];
$stringx = $time.",".$values[32];
$stringy = $time.",".$values[33];
$stringz = $time.",".$values[34];
if ($run4_current >= $run4_count) {
goto skip_4;
                        }
print OUTX $stringx . "\n";
print OUTY $stringy . "\n";
print OUTZ $stringz . "\n";
print OUTCOMBINEDX $stringx . "\n";
print OUTCOMBINEDY $stringy . "\n";
print OUTCOMBINEDZ $stringz . "\n";
skip_4:
         }

                                  }
close INFILE;
close OUTX;
close OUTY;
close OUTZ;

print stdout "all done Run4"."\n";
#################################################################3
##
##
```

```
##
$filelocation = "N:\\Linear\\runs\\prod\\run5\\Linear_21slab_run5.log";
$xdata = ">N:\\Linear\\runs\\prod\\run5\\x_runtime_pressures.txt";
$ydata = ">N:\\Linear\\runs\\prod\\run5\\y_runtime_pressures.txt";
$zdata = ">N:\\Linear\\runs\\prod\\run5\\z_runtime_pressures.txt";
open (INFILE, $filelocation);
open (OUTX, $xdata);
open (OUTY, $ydata);
open (OUTZ, $zdata);
$skip_first = 0;
while (<INFILE>)                  {
@values = split(/ /,$_);
#print stdout ($values[0]);
if ($values[0] eq "PRESSUREPROFILE\:") {
$run5_current++;
#$time = ($values[1]* .0000000000000001);
$time = $values[1];
$stringx = $time.",".$values[32];
$stringy = $time.",".$values[33];
$stringz = $time.",".$values[34];
if ($run5_current >= $run5_count) {
goto skip_5;
                        }
print OUTX $stringx . "\n";
print OUTY $stringy . "\n";
print OUTZ $stringz . "\n";
print OUTCOMBINEDX $stringx . "\n";
print OUTCOMBINEDY $stringy . "\n";
print OUTCOMBINEDZ $stringz . "\n";
skip_5:
          }

                                  }
close INFILE;
close OUTX;
close OUTY;
close OUTZ;
close OUTCOMBINEDX;
close OUTCOMBINEDY;
close OUTCOMBINEDZ;
```

print stdout "all done Run5 and Combined";

Sum_Runtime_Ewald_Pressures_32_33_34.pl (linear)

Sum_Runtime_Ewald_Pressures_11_12_13.pl (parallel)

Sum_Runtime_Ewald_Pressures_14_15_16.pl (perpt)

Sum_Runtime_Ewald_Pressures_11_12_13.pl (skew)

Sum_Runtime_Ewald_Pressures_32_33_34.pl (random)

(This is the final set of a series of programs used to open the NAMD log file outputs from each of the geometric configuration files for experiment #2. Again they are virtually identical to each other. They open the sorted results from the Ewald parse operation and the runtime parse operation and sum them together. The output from these files is the basis for all subsequent

analysis. The exact code from Sum_Runtime_Ewald_Pressures_32_33_34.pl for the original linear system is listed below.)

```
############################################################3
##
## 21 slab cell midway is slab 11
## Corresponding to string position 32 33 34 with time at position 1
##
$xewalddata_run1 = "N:\\Linear\\runs\\prod\\run1\\x_ewald_pressures.txt";
$xruntimedata_run1 = "N:\\Linear\\runs\\prod\\run1\\x_runtime_pressures.txt";
$yewalddata_run1 = "N:\\Linear\\runs\\prod\\run1\\y_ewald_pressures.txt";
$yruntimedata_run1 = "N:\\Linear\\runs\\prod\\run1\\y_runtime_pressures.txt";
$zewalddata_run1 = "N:\\Linear\\runs\\prod\\run1\\z_ewald_pressures.txt";
$zruntimedata_run1 = "N:\\Linear\\runs\\prod\\run1\\z_runtime_pressures.txt";

$xewalddata_run2 = "N:\\Linear\\runs\\prod\\run2\\x_ewald_pressures.txt";
$xruntimedata_run2 = "N:\\Linear\\runs\\prod\\run2\\x_runtime_pressures.txt";
$yewalddata_run2 = "N:\\Linear\\runs\\prod\\run2\\y_ewald_pressures.txt";
$yruntimedata_run2 = "N:\\Linear\\runs\\prod\\run2\\y_runtime_pressures.txt";
$zewalddata_run2 = "N:\\Linear\\runs\\prod\\run2\\z_ewald_pressures.txt";
$zruntimedata_run2 = "N:\\Linear\\runs\\prod\\run2\\z_runtime_pressures.txt";

$xewalddata_run3 = "N:\\Linear\\runs\\prod\\run3\\x_ewald_pressures.txt";
$xruntimedata_run3 = "N:\\Linear\\runs\\prod\\run3\\x_runtime_pressures.txt";
$yewalddata_run3 = "N:\\Linear\\runs\\prod\\run3\\y_ewald_pressures.txt";
$yruntimedata_run3 = "N:\\Linear\\runs\\prod\\run3\\y_runtime_pressures.txt";
$zewalddata_run3 = "N:\\Linear\\runs\\prod\\run3\\z_ewald_pressures.txt";
$zruntimedata_run3 = "N:\\Linear\\runs\\prod\\run3\\z_runtime_pressures.txt";

$xewalddata_run4 = "N:\\Linear\\runs\\prod\\run4\\x_ewald_pressures.txt";
$xruntimedata_run4 = "N:\\Linear\\runs\\prod\\run4\\x_runtime_pressures.txt";
$yewalddata_run4 = "N:\\Linear\\runs\\prod\\run4\\y_ewald_pressures.txt";
$yruntimedata_run4 = "N:\\Linear\\runs\\prod\\run4\\y_runtime_pressures.txt";
$zewalddata_run4 = "N:\\Linear\\runs\\prod\\run4\\z_ewald_pressures.txt";
$zruntimedata_run4 = "N:\\Linear\\runs\\prod\\run4\\z_runtime_pressures.txt";

$xewalddata_run5 = "N:\\Linear\\runs\\prod\\run5\\x_ewald_pressures.txt";
$xruntimedata_run5 = "N:\\Linear\\runs\\prod\\run5\\x_runtime_pressures.txt";
$yewalddata_run5 = "N:\\Linear\\runs\\prod\\run5\\y_ewald_pressures.txt";
$yruntimedata_run5 = "N:\\Linear\\runs\\prod\\run5\\y_runtime_pressures.txt";
$zewalddata_run5 = "N:\\Linear\\runs\\prod\\run5\\z_ewald_pressures.txt";
$zruntimedata_run5 = "N:\\Linear\\runs\\prod\\run5\\z_runtime_pressures.txt";

$xewalddata_combined = "N:\\Linear\\runs\\prod\\run_combined\\x_ewald_pressures.txt";
$xruntimedata_combined = "N:\\Linear\\runs\\prod\\run_combined\\x_runtime_pressures.txt";
$yewalddata_combined = "N:\\Linear\\runs\\prod\\run_combined\\y_ewald_pressures.txt";
$yruntimedata_combined = "N:\\Linear\\runs\\prod\\run_combined\\y_runtime_pressures.txt";
$zewalddata_combined = "N:\\Linear\\runs\\prod\\run_combined\\z_ewald_pressures.txt";
$zruntimedata_combined = "N:\\Linear\\runs\\prod\\run_combined\\z_runtime_pressures.txt";

$sum_xdata_run1 = ">N:\\Linear\\runs\\prod\\run1\\x_sumation_pressures.txt";
$sum_ydata_run1 = ">N:\\Linear\\runs\\prod\\run1\\y_sumation_pressures.txt";
$sum_zdata_run1 = ">N:\\Linear\\runs\\prod\\run1\\z_sumation_pressures.txt";

$sum_xdata_run2 = ">N:\\Linear\\runs\\prod\\run2\\x_sumation_pressures.txt";
$sum_ydata_run2 = ">N:\\Linear\\runs\\prod\\run2\\y_sumation_pressures.txt";
$sum_zdata_run2 = ">N:\\Linear\\runs\\prod\\run2\\z_sumation_pressures.txt";

$sum_xdata_run3 = ">N:\\Linear\\runs\\prod\\run3\\x_sumation_pressures.txt";
$sum_ydata_run3 = ">N:\\Linear\\runs\\prod\\run3\\y_sumation_pressures.txt";
$sum_zdata_run3 = ">N:\\Linear\\runs\\prod\\run3\\z_sumation_pressures.txt";

$sum_xdata_run4 = ">N:\\Linear\\runs\\prod\\run4\\x_sumation_pressures.txt";
$sum_ydata_run4 = ">N:\\Linear\\runs\\prod\\run4\\y_sumation_pressures.txt";
$sum_zdata_run4 = ">N:\\Linear\\runs\\prod\\run4\\z_sumation_pressures.txt";

$sum_xdata_run5 = ">N:\\Linear\\runs\\prod\\run5\\x_sumation_pressures.txt";
$sum_ydata_run5 = ">N:\\Linear\\runs\\prod\\run5\\y_sumation_pressures.txt";
$sum_zdata_run5 = ">N:\\Linear\\runs\\prod\\run5\\z_sumation_pressures.txt";

$sum_xdata_combined = ">N:\\Linear\\runs\\prod\\run_combined\\x_sumation_pressures.txt";
$sum_ydata_combined = ">N:\\Linear\\runs\\prod\\run_combined\\y_sumation_pressures.txt";
$sum_zdata_combined = ">N:\\Linear\\runs\\prod\\run_combined\\z_sumation_pressures.txt";
```

```
print  "Processing Run 1 DATA"."\n";
open (XEINFILE_RUN1, $xewalddata_run1);
#open (YEINFILE, $yewalddata);
#open (ZEINFILE, $zewalddata);
open (XRINFILE_RUN1, $xruntimedata_run1);
#open (YRINFILE, $yruntimedata);
#open (ZRINFILE, $zruntimedata);


open (OUTX, $sum_xdata_run1);
open (OUTY, $sum_ydata_run1);
open (OUTZ, $sum_zdata_run1);
$i = 0;
while (<XEINFILE_RUN1>)                 {
#@values = split(/ /,$_);
$stringe_x[$i] = $_;
$i++;
                                    }
$i = 0;
while (<XRINFILE_RUN1>)           {
$stringr_x[$i] = $_;
$i++;
                                    }
for ($count = 0; $count < ($i-1); $count++)
{
@value1 = split(/,/,$stringe_x[$count]);
@value2 = split(/,/,$stringr_x[$count]);
chomp @value1;
chomp @value2;
$sum_x = ($value1[1] + $value2[1]);
$pstring = $value1[1].",".$value2[1].",".$sum_x."\n";
print OUTX $pstring;
}

close XEINFILE_RUN1;
close XRINFILE_RUN1;


print  "All done X"."\n";

#open (XEINFILE, $xewalddata);
#open (YEINFILE, $yewalddata);
open (YEINFILE_RUN1, $yewalddata_run1);
#open (XRINFILE, $xruntimedata);
#open (YRINFILE, $yruntimedata);
open (YRINFILE_RUN1, $yruntimedata_run1);


#open (OUTX, $sum_xdata);
#open (OUTY, $sum_ydata);
$i = 0;
while (<YEINFILE_RUN1>)                 {
#@values = split(/ /,$_);
$stringe_y[$i] = $_;
$i++;
                                    }
$i = 0;
while (<YRINFILE_RUN1>)           {
$stringr_y[$i] = $_;
$i++;
                                    }
for ($count = 0; $count < ($i-1); $count++)
{
@value1 = split(/,/,$stringe_y[$count]);
@value2 = split(/,/,$stringr_y[$count]);
chomp @value1;
chomp @value2;
$sum_y = ($value1[1] + $value2[1]);
$pstring = $value1[1].",".$value2[1].",".$sum_y."\n";
print OUTY $pstring;
}

close YEINFILE_RUN1;
close YRINFILE_RUN1;
```

```
print  "All done Y"."\n";
#open (XEINFILE, $xewalddata);
#open (YEINFILE, $yewalddata);
open (ZEINFILE_RUN1, $zewalddata_run1);
#open (XRINFILE, $xruntimedata);
#open (YRINFILE, $yruntimedata);
open (ZRINFILE_RUN1, $zruntimedata_run1);


#open (OUTX, $sum_xdata);
#open (OUTY, $sum_ydata);
#open (OUTZ, $sum_zdata);
$i = 0;
while (<ZEINFILE_RUN1>)                 {
#@values = split(/ /,$_);
$stringe_z[$i] = $_;
$i++;
                                  }
$i = 0;
while (<ZRINFILE_RUN1>)           {
$stringr_z[$i] = $_;
$i++;
                                  }
for ($count = 0; $count < ($i-1); $count++)
{
@value1 = split(/,/,$stringe_z[$count]);
@value2 = split(/,/,$stringr_z[$count]);
chomp @value1;
chomp @value2;
$sum_z = ($value1[1] + $value2[1]);
$pstring = $value1[1].",".$value2[1].",".$sum_z."\n";
print OUTZ $pstring;
}

close ZEINFILE_RUN1;
close ZRINFILE_RUN1;

close OUTX;
close OUTY;
close OUTZ;
print  "All done Z"."\n";

#  RUN 2 processing   ###########################################################
 print  "Processing Run 2 DATA"."\n";
open (XEINFILE_RUN2, $xewalddata_run2);
#open (YEINFILE, $yewalddata);
#open (ZEINFILE, $zewalddata);
open (XRINFILE_RUN2, $xruntimedata_run2);
#open (YRINFILE, $yruntimedata);
#open (ZRINFILE, $zruntimedata);


open (OUTX, $sum_xdata_run2);
open (OUTY, $sum_ydata_run2);
open (OUTZ, $sum_zdata_run2);
$i = 0;
while (<XEINFILE_RUN2>)                 {
#@values = split(/ /,$_);
$stringe_x[$i] = $_;
$i++;
                                  }
$i = 0;
while (<XRINFILE_RUN2>)           {
$stringr_x[$i] = $_;
$i++;
                                  }
for ($count = 0; $count < ($i-1); $count++)
{
@value1 = split(/,/,$stringe_x[$count]);
@value2 = split(/,/,$stringr_x[$count]);
chomp @value1;
chomp @value2;
$sum_x = ($value1[1] + $value2[1]);
$pstring = $value1[1].",".$value2[1].",".$sum_x."\n";
print OUTX $pstring;
}
```

```
close XEINFILE_RUN2;
close XRINFILE_RUN2;


print  "All done X"."\n";

#open (XEINFILE, $xewalddata);
#open (YEINFILE, $yewalddata);
open (YEINFILE_RUN2, $yewalddata_run2);
#open (XRINFILE, $xruntimedata);
#open (YRINFILE, $yruntimedata);
open (YRINFILE_RUN2, $yruntimedata_run2);


#open (OUTX, $sum_xdata);
#open (OUTY, $sum_ydata);
$i = 0;
while (<YEINFILE_RUN2>)                 {
#@values = split(/ /,$_);
$stringe_y[$i] = $_;
$i++;
                                  }
$i = 0;
while (<YRINFILE_RUN2>)           {
$stringr_y[$i] = $_;
$i++;
                                  }
for ($count = 0; $count < ($i-1); $count++)
{
@value1 = split(/,/,$stringe_y[$count]);
@value2 = split(/,/,$stringr_y[$count]);
chomp @value1;
chomp @value2;
$sum_y = ($value1[1] + $value2[1]);
$pstring = $value1[1].",".$value2[1].",".$sum_y."\n";
print OUTY $pstring;
}

close YEINFILE_RUN2;
close YRINFILE_RUN2;

print  "All done Y"."\n";
#open (XEINFILE, $xewalddata);
#open (YEINFILE, $yewalddata);
open (ZEINFILE_RUN2, $zewalddata_run2);
#open (XRINFILE, $xruntimedata);
#open (YRINFILE, $yruntimedata);
open (ZRINFILE_RUN2, $zruntimedata_run2);


#open (OUTX, $sum_xdata);
#open (OUTY, $sum_ydata);
#open (OUTZ, $sum_zdata);
$i = 0;
while (<ZEINFILE_RUN2>)                 {
#@values = split(/ /,$_);
$stringe_z[$i] = $_;
$i++;
                                  }
$i = 0;
while (<ZRINFILE_RUN2>)           {
$stringr_z[$i] = $_;
$i++;
                                  }
for ($count = 0; $count < ($i-1); $count++)
{
@value1 = split(/,/,$stringe_z[$count]);
@value2 = split(/,/,$stringr_z[$count]);
chomp @value1;
chomp @value2;
$sum_z = ($value1[1] + $value2[1]);
$pstring = $value1[1].",".$value2[1].",".$sum_z."\n";
print OUTZ $pstring;
}

close ZEINFILE_RUN2;
close ZRINFILE_RUN2;
```

```
close OUTX;
close OUTY;
close OUTZ;
print  "All done Z"."\n";
#  RUN 3 processing   ###########################################################
print  "Processing Run 3 DATA"."\n";
open (XEINFILE_RUN3, $xewalddata_run3);
#open (YEINFILE, $yewalddata);
#open (ZEINFILE, $zewalddata);
open (XRINFILE_RUN3, $xruntimedata_run3);
#open (YRINFILE, $yruntimedata);
#open (ZRINFILE, $zruntimedata);


open (OUTX, $sum_xdata_run3);
open (OUTY, $sum_ydata_run3);
open (OUTZ, $sum_zdata_run3);
$i = 0;
while (<XEINFILE_RUN3>)                 {
#@values = split(/ /,$_);
$stringe_x[$i] = $_;
$i++;
                                  }
$i = 0;
while (<XRINFILE_RUN3>)           {
$stringr_x[$i] = $_;
$i++;
                                  }
for ($count = 0; $count < ($i-1); $count++)
{
@value1 = split(/,/,$stringe_x[$count]);
@value2 = split(/,/,$stringr_x[$count]);
chomp @value1;
chomp @value2;
$sum_x = ($value1[1] + $value2[1]);
$pstring = $value1[1].",".$value2[1].",".$sum_x."\n";
print OUTX $pstring;
}

close XEINFILE_RUN3;
close XRINFILE_RUN3;


print  "All done X"."\n";

#open (XEINFILE, $xewalddata);
#open (YEINFILE, $yewalddata);
open (YEINFILE_RUN3, $yewalddata_run3);
#open (XRINFILE, $xruntimedata);
#open (YRINFILE, $yruntimedata);
open (YRINFILE_RUN3, $yruntimedata_run3);


#open (OUTX, $sum_xdata);
#open (OUTY, $sum_ydata);
$i = 0;
while (<YEINFILE_RUN3>)                 {
#@values = split(/ /,$_);
$stringe_y[$i] = $_;
$i++;
                                  }
$i = 0;
while (<YRINFILE_RUN3>)           {
$stringr_y[$i] = $_;
$i++;
                                  }
for ($count = 0; $count < ($i-1); $count++)
{
@value1 = split(/,/,$stringe_y[$count]);
@value2 = split(/,/,$stringr_y[$count]);
chomp @value1;
chomp @value2;
$sum_y = ($value1[1] + $value2[1]);
$pstring = $value1[1].",".$value2[1].",".$sum_y."\n";
print OUTY $pstring;
}
```

```
close YEINFILE_RUN3;
close YRINFILE_RUN3;

print  "All done Y"."\n";
#open (XEINFILE, $xewalddata);
#open (YEINFILE, $yewalddata);
open (ZEINFILE_RUN3, $zewalddata_run3);
#open (XRINFILE, $xruntimedata);
#open (YRINFILE, $yruntimedata);
open (ZRINFILE_RUN3, $zruntimedata_run3);


#open (OUTX, $sum_xdata);
#open (OUTY, $sum_ydata);
#open (OUTZ, $sum_zdata);
$i = 0;
while (<ZEINFILE_RUN3>)                 {
#@values = split(/ /,$_);
$stringe_z[$i] = $_;
$i++;
                                  }
$i = 0;
while (<ZRINFILE_RUN3>)           {
$stringr_z[$i] = $_;
$i++;
                                  }
for ($count = 0; $count < ($i-1); $count++)
{
@value1 = split(/,/,$stringe_z[$count]);
@value2 = split(/,/,$stringr_z[$count]);
chomp @value1;
chomp @value2;
$sum_z = ($value1[1] + $value2[1]);
$pstring = $value1[1].",".$value2[1].",".$sum_z."\n";
print OUTZ $pstring;
}

close ZEINFILE_RUN3;
close ZRINFILE_RUN3;

close OUTX;
close OUTY;
close OUTZ;
print  "All done Z"."\n";
#  RUN 4 processing   ###############################################################
print  "Processing Run 4 DATA"."\n" ;
open (XEINFILE_RUN4, $xewalddata_run4);
#open (YEINFILE, $yewalddata);
#open (ZEINFILE, $zewalddata);
open (XRINFILE_RUN4, $xruntimedata_run4);
#open (YRINFILE, $yruntimedata);
#open (ZRINFILE, $zruntimedata);


open (OUTX, $sum_xdata_run4);
open (OUTY, $sum_ydata_run4);
open (OUTZ, $sum_zdata_run4);
$i = 0;
while (<XEINFILE_RUN4>)                 {
#@values = split(/ /,$_);
$stringe_x[$i] = $_;
$i++;
                                  }
$i = 0;
while (<XRINFILE_RUN4>)           {
$stringr_x[$i] = $_;
$i++;
                                  }
for ($count = 0; $count < ($i-1); $count++)
{
@value1 = split(/,/,$stringe_x[$count]);
@value2 = split(/,/,$stringr_x[$count]);
chomp @value1;
chomp @value2;
$sum_x = ($value1[1] + $value2[1]);
$pstring = $value1[1].",".$value2[1].",".$sum_x."\n";
```

```
print OUTX $pstring;
}

close XEINFILE_RUN4;
close XRINFILE_RUN4;


print  "All done X"."\n";

#open (XEINFILE, $xewalddata);
#open (YEINFILE, $yewalddata);
open (YEINFILE_RUN4, $yewalddata_run4);
#open (XRINFILE, $xruntimedata);
#open (YRINFILE, $yruntimedata);
open (YRINFILE_RUN4, $yruntimedata_run4);


#open (OUTX, $sum_xdata);
#open (OUTY, $sum_ydata);
$i = 0;
while (<YEINFILE_RUN4>)                  {
#@values = split(/ /,$_);
$stringe_y[$i] = $_;
$i++;
                                    }
$i = 0;
while (<YRINFILE_RUN4>)           {
$stringr_y[$i] = $_;
$i++;
                                    }
for ($count = 0; $count < ($i-1); $count++)
{
@value1 = split(/,/,$stringe_y[$count]);
@value2 = split(/,/,$stringr_y[$count]);
chomp @value1;
chomp @value2;
$sum_y = ($value1[1] + $value2[1]);
$pstring = $value1[1].",".$value2[1].",".$sum_y."\n";
print OUTY $pstring;
}

close YEINFILE_RUN4;
close YRINFILE_RUN4;

print  "All done Y"."\n";
#open (XEINFILE, $xewalddata);
#open (YEINFILE, $yewalddata);
open (ZEINFILE_RUN4, $zewalddata_run4);
#open (XRINFILE, $xruntimedata);
#open (YRINFILE, $yruntimedata);
open (ZRINFILE_RUN4, $zruntimedata_run4);


#open (OUTX, $sum_xdata);
#open (OUTY, $sum_ydata);
#open (OUTZ, $sum_zdata);
$i = 0;
while (<ZEINFILE_RUN4>)                  {
#@values = split(/ /,$_);
$stringe_z[$i] = $_;
$i++;
                                    }
$i = 0;
while (<ZRINFILE_RUN4>)           {
$stringr_z[$i] = $_;
$i++;
                                    }
for ($count = 0; $count < ($i-1); $count++)
{
@value1 = split(/,/,$stringe_z[$count]);
@value2 = split(/,/,$stringr_z[$count]);
chomp @value1;
chomp @value2;
$sum_z = ($value1[1] + $value2[1]);
$pstring = $value1[1].",".$value2[1].",".$sum_z."\n";
print OUTZ $pstring;
}
```

```
close ZEINFILE_RUN4;
close ZRINFILE_RUN4;

close OUTX;
close OUTY;
close OUTZ;
print  "All done Z"."\n";
#  RUN 5 processing   ############################################################
print  "Processing Run 5 DATA"."\n";
open (XEINFILE_RUN5, $xewalddata_run5);
#open (YEINFILE, $yewalddata);
#open (ZEINFILE, $zewalddata);
open (XRINFILE_RUN5, $xruntimedata_run5);
#open (YRINFILE, $yruntimedata);
#open (ZRINFILE, $zruntimedata);


open (OUTX, $sum_xdata_run5);
open (OUTY, $sum_ydata_run5);
open (OUTZ, $sum_zdata_run5);
$i = 0;
while (<XEINFILE_RUN5>)                 {
#@values = split(/ /,$_);
$stringe_x[$i] = $_;
$i++;
                                  }
$i = 0;
while (<XRINFILE_RUN5>)           {
$stringr_x[$i] = $_;
$i++;
                                  }
for ($count = 0; $count < ($i-1); $count++)
{
@value1 = split(/,/,$stringe_x[$count]);
@value2 = split(/,/,$stringr_x[$count]);
chomp @value1;
chomp @value2;
$sum_x = ($value1[1] + $value2[1]);
$pstring = $value1[1].",".$value2[1].",".$sum_x."\n";
print OUTX $pstring;
}

close XEINFILE_RUN5;
close XRINFILE_RUN5;


print  "All done X"."\n";

#open (XEINFILE, $xewalddata);
#open (YEINFILE, $yewalddata);
open (YEINFILE_RUN5, $yewalddata_run5);
#open (XRINFILE, $xruntimedata);
#open (YRINFILE, $yruntimedata);
open (YRINFILE_RUN5, $yruntimedata_run5);


#open (OUTX, $sum_xdata);
#open (OUTY, $sum_ydata);
$i = 0;
while (<YEINFILE_RUN5>)                 {
#@values = split(/ /,$_);
$stringe_y[$i] = $_;
$i++;
                                  }
$i = 0;
while (<YRINFILE_RUN5>)           {
$stringr_y[$i] = $_;
$i++;
                                  }
for ($count = 0; $count < ($i-1); $count++)
{
@value1 = split(/,/,$stringe_y[$count]);
@value2 = split(/,/,$stringr_y[$count]);
chomp @value1;
chomp @value2;
$sum_y = ($value1[1] + $value2[1]);
```

```
$pstring = $value1[1].",".$value2[1].",".$sum_y."\n";
print OUTY $pstring;
}

close YEINFILE_RUN5;
close YRINFILE_RUN5;

print  "All done Y"."\n";
#open (XEINFILE, $xewalddata);
#open (YEINFILE, $yewalddata);
open (ZEINFILE_RUN5, $zewalddata_run5);
#open (XRINFILE, $xruntimedata);
#open (YRINFILE, $yruntimedata);
open (ZRINFILE_RUN5, $zruntimedata_run5);


#open (OUTX, $sum_xdata);
#open (OUTY, $sum_ydata);
#open (OUTZ, $sum_zdata);
$i = 0;
while (<ZEINFILE_RUN5>)                  {
#@values = split(/ /,$_);
$stringe_z[$i] = $_;
$i++;
                                  }
$i = 0;
while (<ZRINFILE_RUN5>)           {
$stringr_z[$i] = $_;
$i++;
                                  }
for ($count = 0; $count < ($i-1); $count++)
{
@value1 = split(/,/,$stringe_z[$count]);
@value2 = split(/,/,$stringr_z[$count]);
chomp @value1;
chomp @value2;
$sum_z = ($value1[1] + $value2[1]);
$pstring = $value1[1].",".$value2[1].",".$sum_z."\n";
print OUTZ $pstring;
}

close ZEINFILE_RUN5;
close ZRINFILE_RUN5;

close OUTX;
close OUTY;
close OUTZ;
print  "All done Z"."\n";
#  Combined RUNS processing   ############################################################
print  "Processing COMBINED RUNS DATA"."\n";
open (XEINFILE_COMBINED, $xewalddata_combined);
#open (YEINFILE, $yewalddata);
#open (ZEINFILE, $zewalddata);
open (XRINFILE_COMBINED, $xruntimedata_combined);
#open (YRINFILE, $yruntimedata);
#open (ZRINFILE, $zruntimedata);


open (OUTX, $sum_xdata_combined);
open (OUTY, $sum_ydata_combined);
open (OUTZ, $sum_zdata_combined);
$i = 0;
while (<XEINFILE_COMBINED>)                 {
#@values = split(/ /,$_);
$stringe_x[$i] = $_;
$i++;
                                  }
$i = 0;
while (<XRINFILE_COMBINED>)               {
$stringr_x[$i] = $_;
$i++;
                                  }
for ($count = 0; $count < ($i-1); $count++)
{
@value1 = split(/,/,$stringe_x[$count]);
@value2 = split(/,/,$stringr_x[$count]);
chomp @value1;
```

```
chomp @value2;
$sum_x = ($value1[1] + $value2[1]);
$pstring = $value1[1].",".$value2[1].",".$sum_x."\n";
print OUTX $pstring;
}

close XEINFILE_COMBINED;
close XRINFILE_COMBINED;


print  "All done X"."\n";

#open (XEINFILE, $xewalddata);
#open (YEINFILE, $yewalddata);
open (YEINFILE_COMBINED, $yewalddata_combined);
#open (XRINFILE, $xruntimedata);
#open (YRINFILE, $yruntimedata);
open (YRINFILE_COMBINED, $yruntimedata_combined);


#open (OUTX, $sum_xdata);
#open (OUTY, $sum_ydata);
$i = 0;
while (<YEINFILE_COMBINED>)                  {
#@values = split(/ /,$_);
$stringe_y[$i] = $_;
$i++;
                                    }
$i = 0;
while (<YRINFILE_COMBINED>)                {
$stringr_y[$i] = $_;
$i++;
                                    }
for ($count = 0; $count < ($i-1); $count++)
{
@value1 = split(/,/,$stringe_y[$count]);
@value2 = split(/,/,$stringr_y[$count]);
chomp @value1;
chomp @value2;
$sum_y = ($value1[1] + $value2[1]);
$pstring = $value1[1].",".$value2[1].",".$sum_y."\n";
print OUTY $pstring;
}

close YEINFILE_COMBINED;
close YRINFILE_COMBINED;

print  "All done Y"."\n";
#open (XEINFILE, $xewalddata);
#open (YEINFILE, $yewalddata);
open (ZEINFILE_COMBINED, $zewalddata_combined);
#open (XRINFILE, $xruntimedata);
#open (YRINFILE, $yruntimedata);
open (ZRINFILE_COMBINED, $zruntimedata_combined);


#open (OUTX, $sum_xdata);
#open (OUTY, $sum_ydata);
#open (OUTZ, $sum_zdata);
$i = 0;
while (<ZEINFILE_COMBINED>)                  {
#@values = split(/ /,$_);
$stringe_z[$i] = $_;
$i++;
                                    }
$i = 0;
while (<ZRINFILE_COMBINED>)                {
$stringr_z[$i] = $_;
$i++;
                                    }
for ($count = 0; $count < ($i-1); $count++)
{
@value1 = split(/,/,$stringe_z[$count]);
@value2 = split(/,/,$stringr_z[$count]);
chomp @value1;
chomp @value2;
$sum_z = ($value1[1] + $value2[1]);
```

```
$pstring = $value1[1].",".$value2[1].",".$sum_z."\n";
print OUTZ $pstring;
}

close ZEINFILE_COMBINED;
close ZRINFILE_COMBINED;

close OUTX;
close OUTY;
close OUTZ;
print  "All done Z"."\n";
print  "All done with EVERYTHING"."\n";
```

Match_z_freqs_4systems.pl

(This program was used to open the spectral content result files from 4 MATLAB FFT

Transforms and sort them looking for exact matches.)

```
############################################################3
##
## 4 System Z matches looks for matches between Linear, Parallel, Perpt and Skew
## in the significant z frequency data
## it prints each match into 4system_z_matches.txt output file

$significant_linear_z =   "N:\\Linear\\Analysis\\LinearColumn01z_data.txt";
$significant_parallel_z = "N:\\Parallel\\Analysis\\ParallelColumn01z_data.txt";
$significant_perpt_z =    "N:\\PerpT\\Analysis\\PerpTColumn01z_data.txt";
$significant_skew_z =     "N:\\Skew\\Analysis\\SkewColumn01z_data.txt";
#$significant_linearrandom_z = "N:\\Linearrandom\\Analysis\\LinearrandomColumn01z_data.txt";
#$significant_random_z =  "N:\\Random\\Analysis\\RandomColumn01z_data.txt";

print "Processing"."\n";
$four_system_matches=0;
$five_system_matches=0;
$six_system_matches=0;
$z_matches = ">N:\\summary\\4system_z_matches.txt";
#$five_z_matches = ">N:\\summary\\5system_z_matches.txt";
#$six_z_matches = ">N:\\summary\\6system_z_matches.txt";
open (OUT_Z, $z_matches);
#open (OUT_FIVE_Z, $five_z_matches);
#open (OUT_SIX_Z, $six_z_matches);

open (LINEARINFILE, $significant_linear_z);
$lineartcount=0;
while (<LINEARINFILE>)                                                                  {  # Linear input
loop
$linearcount++;
@linear_values = split(/,/,$_);
chomp @linear_values;
open (SKEWINFILE, $significant_skew_z);
$skewcount=0;
while (<SKEWINFILE>)                                                    { # Skew input loop
$skewcount++;
@skew_values = split(/,/,$_);
chomp @skew_values;
if ($linear_values[0] == $skew_values[0])                               { # linear and skew has matched
print "linear skew match"."\n";
open (PERPTINFILE, $significant_perpt_z);
$perptcount=0;
while (<PERPTINFILE>)                                                   { #Perpt input loop
$perptcount++;
@perpt_values = split(/,/,$_);
chomp @perpt_values;
if ($linear_values[0] == $perpt_values[0])                  { #linear skew perpt matched
print "linear perpt match"."\n";
open (PARALLELINFILE, $significant_parallel_z);
while (<PARALLELINFILE>)                    { #Parallelinfile loop
$parallelcount++;
@parallel_values = split(/,/,$_);
chomp @parallel_values;
if ($linear_values[0] == $parallel_values[0])       { #4 SYSTEM MATCH
```

```
$four_system_matches++;
print "linear parallel (4 system) match"."\n";
print OUT_Z
($linear_values[0].",".$linear_values[1].",".$linear_values[2].",".$linear_values[3].",".$linear_values[4].",".
$linear_values[5].",".$linear_values[6].",".$linear_values[7].",".$linear_values[8]);
print OUT_Z
($parallel_values[0].",".$parallel_values[1].",".$parallel_values[2].",".$parallel_values[3].",".$parallel_valu
es[4].",".$parallel_values[5].",".$parallel_values[6].",".$parallel_values[7].",".$parallel_values[8]);
print OUT_Z
($perpt_values[0].",".$perpt_values[1].",".$perpt_values[2].",".$perpt_values[3].",".$perpt_values[4].",".$perp
t_values[5].",".$perpt_values[6].",".$perpt_values[7].",".$perpt_values[8]);
print OUT_Z
($skew_values[0].",".$skew_values[1].",".$skew_values[2].",".$skew_values[3].",".$skew_values[4].",".$skew_valu
es[5].",".$skew_values[6].",".$skew_values[7].",".$skew_values[8]."\n");
#open (LINEARRANDOMINFILE, $significant_linearrandom_z);
#while (<LINEARRANDOMINFILE>)                                                    { #Linearrandom loop
#@linearrandom_values = split(/,/,$_);
#chomp @linearrandom_values;
#if ($linear_values[0] == $linearrandom_values[0])  { #5 SYSTEM MATCH
#$five_system_matches++;
#print "linear linearrandom (5 system) match"."\n";
#print OUT_FIVE_Z ($linear_values[0].",".$linear_values[1].",".$linear_values[2].",");
#print OUT_FIVE_Z ($parallel_values[0].",".$parallel_values[1].",".$parallel_values[2].",");
#print OUT_FIVE_Z ($perpt_values[0].",".$perpt_values[1].",".$perpt_values[2].",");
#print OUT_FIVE_Z ($skew_values[0].",".$skew_values[1].",".$skew_values[2].",");
#print OUT_FIVE_Z ($linearrandom_values[0].",".$linearrandom_values[1].",".$linearrandom_values[2]."\n");
#open (RANDOMINFILE, $significant_random_z);
#while (<RANDOMINFILE>)  {# random loop
#@random_values = split(/,/,$_);
#chomp @random_values;
#if ($linear_values[0] == $random_values[0])       { #6 SYSTEM MATCH
#$six_system_matches++;
#print "linear random (6 system) match"."\n";
#print OUT_SIX_Z ($linear_values[0].",".$linear_values[1].",".$linear_values[2].",");
#print OUT_SIX_Z ($parallel_values[0].",".$parallel_values[1].",".$parallel_values[2].",");
#print OUT_SIX_Z ($perpt_values[0].",".$perpt_values[1].",".$perpt_values[2].",");
#print OUT_SIX_Z ($skew_values[0].",".$skew_values[1].",".$skew_values[2].",");
#print OUT_SIX_Z ($linearrandom_values[0].",".$linearrandom_values[1].",".$linearrandom_values[2].",");
#print OUT_SIX_Z ($random_values[0].",".$random_values[1].",".$random_values[2]."\n");
#                                     } #6 SYSTEM MATCH
#                        }# random loop
#close RANDOMINFILE;
#                                                  } #5 SYSTEM MATCH
#                                                                    } #Linearrandom loop
#close LINEARRANDOMINFILE;
                                               } #4 SYSTEM MATCH
                                          } #Parallelinfile loop
close PARALLELINFILE;
                                                  }  #linear skew perpt matched
                                                         } #Perpt input loop
close PERPTINFILE;
                                                                  } # linear and skew has matched
                                                                          } # Skew input loop
close SKEWINFILE;
                                                                                   }   # Linear
input loop
close LINEARINFILE;
print "Total Linear Coeffs:  ".$linearcount."  Total 4 system matches:  ".$four_system_matches." Five System
matches: ".$five_system_matches." Six System matches: ".$six_system_matches."\n";
close OUT_Z;
#close OUT_SIX_Z;
#close OUT_FIVE_Z;
print  "All done with EVERYTHING"."\n";
```

Match_z_freqs_4systems_Random.pl

(This program was used to open the spectral content result files from 4 MATLAB FFT Transforms and sort them looking for exact matches, it was modified to open the results from the linear configuration random sequence simulation.)

```
#############################################################3
##
## 4 System Z matches looks for matches between Linear, Parallel, Perpt and Skew
## in the significant z frequency data
## it prints each match into 4system_z_matches.txt output file

$significant_linear_z =   "N:\\Random\\Analysis\\RandomColumn01z_data.txt";
$significant_parallel_z = "N:\\Parallel\\Analysis\\ParallelColumn01z_data.txt";
$significant_perpt_z =    "N:\\PerpT\\Analysis\\PerpTColumn01z_data.txt";
$significant_skew_z =     "N:\\Skew\\Analysis\\SkewColumn01z_data.txt";
#$significant_linearrandom_z = "N:\\Linearrandom\\Analysis\\LinearrandomColumn01z_data.txt";
#$significant_random_z =  "N:\\Random\\Analysis\\RandomColumn01z_data.txt";

print "Processing"."\n";
$four_system_matches=0;
$five_system_matches=0;
$six_system_matches=0;
$z_matches = ">N:\\summary\\4system_z_matches_random.txt";
#$five_z_matches = ">N:\\summary\\5system_z_matches.txt";
#$six_z_matches = ">N:\\summary\\6system_z_matches.txt";
open (OUT_Z, $z_matches);
#open (OUT_FIVE_Z, $five_z_matches);
#open (OUT_SIX_Z, $six_z_matches);

open (LINEARINFILE, $significant_linear_z);
$lineartcount=0;
while (<LINEARINFILE>)                                                                  {  # Linear input
loop
$linearcount++;
@linear_values = split(/,/,$_);
chomp @linear_values;
open (SKEWINFILE, $significant_skew_z);
$skewcount=0;
while (<SKEWINFILE>)                                                        { # Skew input loop
$skewcount++;
@skew_values = split(/,/,$_);
chomp @skew_values;
if ($linear_values[0] == $skew_values[0])                           { # linear and skew has matched
print "linear skew match"."\n";
open (PERPTINFILE, $significant_perpt_z);
$perptcount=0;
while (<PERPTINFILE>)                                               { #Perpt input loop
$perptcount++;
@perpt_values = split(/,/,$_);
chomp @perpt_values;
if ($linear_values[0] == $perpt_values[0])                  { #linear skew perpt matched
print "linear perpt match"."\n";
open (PARALLELINFILE, $significant_parallel_z);
while (<PARALLELINFILE>)                    { #Parallelinfile loop
$parallelcount++;
@parallel_values = split(/,/,$_);
chomp @parallel_values;
if ($linear_values[0] == $parallel_values[0])       { #4 SYSTEM MATCH
$four_system_matches++;
print "linear parallel (4 system) match"."\n";
print OUT_Z
($linear_values[0].",".$linear_values[1].",".$linear_values[2].",".$linear_values[3].",".$linear_values[4].",".
$linear_values[5].",".$linear_values[6].",".$linear_values[7].",".$linear_values[8]);
print OUT_Z
($parallel_values[0].",".$parallel_values[1].",".$parallel_values[2].",".$parallel_values[3].",".$parallel_valu
es[4].",".$parallel_values[5].",".$parallel_values[6].",".$parallel_values[7].",".$parallel_values[8]);
```

```
print OUT_Z
($perpt_values[0].",".$perpt_values[1].",".$perpt_values[2].",".$perpt_values[3].",".$perpt_values[4].",".$perp
t_values[5].",".$perpt_values[6].",".$perpt_values[7].",".$perpt_values[8]);
print OUT_Z
($skew_values[0].",".$skew_values[1].",".$skew_values[2].",".$skew_values[3].",".$skew_values[4].",".$skew_valu
es[5].",".$skew_values[6].",".$skew_values[7].",".$skew_values[8]."\n");
#open (LINEARRANDOMINFILE, $significant_linearrandom_z);
#while (<LINEARRANDOMINFILE>)                                                    { #Linearrandom loop
#@linearrandom_values = split(/,/,$_);
#chomp @linearrandom_values;
#if ($linear_values[0] == $linearrandom_values[0]) { #5 SYSTEM MATCH
#$five_system_matches++;
#print "linear linearrandom (5 system) match"."\n";
#print OUT_FIVE_Z ($linear_values[0].",".$linear_values[1].",".$linear_values[2].",");
#print OUT_FIVE_Z ($parallel_values[0].",".$parallel_values[1].",".$parallel_values[2].",");
#print OUT_FIVE_Z ($perpt_values[0].",".$perpt_values[1].",".$perpt_values[2].",");
#print OUT_FIVE_Z ($skew_values[0].",".$skew_values[1].",".$skew_values[2].",");
#print OUT_FIVE_Z ($linearrandom_values[0].",".$linearrandom_values[1].",".$linearrandom_values[2]."\n");
#open (RANDOMINFILE, $significant_random_z);
#while (<RANDOMINFILE>)  {# random loop
#@random_values = split(/,/,$_);
#chomp @random_values;
#if ($linear_values[0] == $random_values[0])       { #6 SYSTEM MATCH
#$six_system_matches++;
#print "linear random (6 system) match"."\n";
#print OUT_SIX_Z ($linear_values[0].",".$linear_values[1].",".$linear_values[2].",");
#print OUT_SIX_Z ($parallel_values[0].",".$parallel_values[1].",".$parallel_values[2].",");
#print OUT_SIX_Z ($perpt_values[0].",".$perpt_values[1].",".$perpt_values[2].",");
#print OUT_SIX_Z ($skew_values[0].",".$skew_values[1].",".$skew_values[2].",");
#print OUT_SIX_Z ($linearrandom_values[0].",".$linearrandom_values[1].",".$linearrandom_values[2].",");
#print OUT_SIX_Z ($random_values[0].",".$random_values[1].",".$random_values[2]."\n");
#                                       } #6 SYSTEM MATCH
#                       }# random loop
#close RANDOMINFILE;
#                                                   } #5 SYSTEM MATCH
#                                                                   } #Linearrandom loop
#close LINEARRANDOMINFILE;
                                                  } #4 SYSTEM MATCH
                                               } #Parallelinfile loop
close PARALLELINFILE;
                                                   }  #linear skew perpt matched
                                                           } #Perpt input loop
close PERPTINFILE;
                                                                           } # linear and skew has matched
                                                                                   } # Skew input loop
close SKEWINFILE;
                                                                                            }   # Linear
input loop
close LINEARINFILE;
print "Total Linear Coeffs:  ".$linearcount."  Total 4 system matches:  ".$four_system_matches." Five System
matches: ".$five_system_matches." Six System matches: ".$six_system_matches."\n";
close OUT_Z;
#close OUT_SIX_Z;
#close OUT_FIVE_Z;
print  "All done with EVERYTHING"."\n";
```

APPENDIX-F: PRIMARY MATLAB PROGRAMS

Hartley_Combo_Final_Rev5.m

(This program is the main statistical analysis tool. It was run on every system in the project with changes made only to the file names.)

clear all

```
%%This is where the actual data is normally read in
%file_name='Column16_sig'; %This is a variable that prints on graphs and reports
%file_name='LinearColumn16decx'; %This is a variable that prints on graphs and reports
%file_name='LinearColumn16y'; %This is a variable that prints on graphs
%file_name='LinearColumn01x'; %This is for no decimation 99999 points
%file_name='LinearColumn16decx'; %This is for no decimation
%file_name='LinearColumn16z'; %This is a variable that prints on graphs and reports
%file_name='x_sumation_pressures'; %This is a variable that prints on
%graphs and reports
file_name='LinearColumn01z'; %
data_type=1;% 1--for Single     2--for Decimated      3--for Windowed  4 -- Resultant Pressures
if (data_type==1)
window_data=0;
sample_size=16;
decimation_step=1;
dec_data=0;
single_data_set=1;
data_set_length=99993;
end
if (data_type==2)
window_data=0;
sample_size=16;
decimation_step=16;
dec_data=1;
single_data_set=0;
end
if (data_type==3)
window_data=1;
sample_size=16;
decimation_step=1;
dec_data=0;
single_data_set=0;
end
if (data_type==4)
window_data=0;
sample_size=16;
decimation_step=1;
dec_data=0;
single_data_set=1;
data_set_length=99990;
end

%%window_data=1; %Set to 1 if data windowed
%%sample_size=16; %Set this to 16 for 16 sample windows or 16 step decimations used in SEM calculations
%%decimation_step=1; %Set to 1 if data not decimated *IMPORTANT* >>used in Tint calc!
%%dec_data=1; %Set to 1 if data decimated
%%single_data_set=0; %Set to 1 if data not windowed or decimated
spr=10; %pressureprofile output was 10 steps per cycle during simulation
top_harmonics=5; %This is desired number to return to time domain   %return_harmonics needs to be odd
v=num2str(top_harmonics);%This sets v equal to a string that can be printed
alpha = 0.001; %This is alpha for all stat tests
alpha_string=num2str(alpha);%This sets alpha_string equal to a string that can be printed


file_w_ext=[file_name '.txt'];
load_string=['./' (file_w_ext)];


if (data_type==1)
%clip data to 99995 points
whole_data= load(load_string);
  for clip=1:data_set_length
  file_data(clip)=whole_data(clip);
```

```
  end
end
if (data_type==2)
%don't do anything
file_data= load(load_string);
end
if (data_type==3)
%don't do anything
file_data= load(load_string);
end
if (data_type==4)
whole_data= load(load_string);
  for clip=1:data_set_length
  file_data(clip)=whole_data(clip);
  end
end
%pressure=Column16(1,:);
if (dec_data==1)  %Testing for non-decimated data
pressure=mean(file_data);% This takes the average of decimated columns for determiniation of later confidence
bands on graph
else
    if (window_data==1)
    pressure=mean(file_data);
    else
        pressure=file_data; % This is for non-decimated non-windowed data
    end
end
%load ../runs/prod/run_combined/x_sumation_pressures.txt
%load ../runs/prod/run_combined/y_sumation_pressures.txt
%load ../runs/prod/run_combined/z_sumation_pressures.txt
%pressure=x_sumation_pressures_small(:,3);
%pressurey=y_sumation_pressures(:,3);
%pressurez=z_sumation_pressures(:,3);
n = length(pressure);%%%%%n = 20000; %number of samples or simulation run steps
if mod(n,2)
   %disp('odd')
else
   %disp('even')
   pressure(n+1)=pressure(n); %Make sure data set always odd
   n=n+1;
end

%if (dec_data==1)
%Tint = 2e-15*sample_adjust*spr;  %this should be 2fs times 10 steps per output for profile pressure Sample
Time
%else
%Tint = 2e-15*sample_adjust*spr; %This applies for windowed and serial data
%end

%This applies to all data
Tint = 2e-15*decimation_step*spr;  %this should be 2fs times 10 steps per output for profile pressure Sample
Time

ftop = (1/2)*1/Tint; %simple nyquist frequency
fbottom = 1/(n*Tint); %1 wave over the window
SampTimeActual=n*Tint;%%%%%SampTimeActual=20000*Tint;
SampTime = Tint*n; %Total time intervalx
Fsamp=1/Tint; %sampling frequency is 1/interval samples every second in Hz or sampling rate
%take fft of data
%divide by number of data points Dme
%take abs of fft and square it
%subtract the first coeficient returned by fft wich is the DC term
%%%%%%%%%%%%%%%%%Verification with Sample DATA is done here%%%%%%%%%%%%
%pressure=[0.7712
%   -2.1036
%    1.1951
%    1.8159
%    0.7476
%    1.1402
%    0.4931
%    0.5502
%    0.2417
%    0.0489
%   -2.1952];
%n=length(pressure);
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%  FFT    %%%%%%%%%%%%%%%%%
raw_coeffs=fft(pressure);
```

```
Fmax=length(raw_coeffs);
% F(-n)= F(n+1)+oppsite_factor
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%  END FFT

%Standardize Coefficeints
%Subtracting off mean and dividing by Standard deviation
mean_raw_coeffs=mean(raw_coeffs);
stdev=std(raw_coeffs);
raw_coeffs_standardized=(raw_coeffs-mean_raw_coeffs)/stdev;%This needed to make coefficients standardized
Gaussian random variable with zero
%raw_coeffs_standardized=raw_coeffs/n;%to Match Homework9 from Thibos

%Calculate the Pos Raw Coeffs Stndrdzed
y=1;
while (y<floor(n/2)+1)
Pos_raw_coeffs_standardized(y)=raw_coeffs_standardized(y);
y=y+1;
end

%mean and unit variance.  Squaring them then means they will then be
%distrubuted as Chi-squared.  this will be useful when doing confidence
%intervals below
%End standardizeing Coefficients
%Also Calculate the

%Need to square coefficients to get power
p=raw_coeffs_standardized;                    % (took n out)This matches homework 9 don't know if its
standardized or not
pwr=abs(p).^2;             %pwr is power of standardized coeffients freq goes 0 to midway
TotPwr=sum(pwr)-pwr(1);    %This is actually twice the power with DC subtracted
Pos_TotPwr=TotPwr/2;           %Divide by 2 to get power of 1 side

%for u=2:(length(pwr))    %This works because pwr is DC f1 f2 f3 -f3 -f2 -f1
%pwr_no_dc(u-1)=pwr(u);       %This is power series without DC
%end
pwr_dc_zero=pwr;
pwr_dc_zero(1)=0;
pwr_no_dc=pwr(2:floor(n/2)+1);
pos_pwr=pwr(1:(floor(n/2)+1));%%%%Positive half of spectrum only with DC in first position
pos_pwr_no_dc=pwr(2:(floor(n/2)+1));%%%%Positive half of spectrum only with f1 in first position
pos_pwr_dc_zero=pwr_dc_zero(1:(floor(n/2)+1)); %positive have of spectrum with DC at zero in first position
freq = [0:(floor(n/2))]/(SampTime); %find the corresponding frequency in Hz This assumes shifted coeficients
freq_no_dc = [1:(floor(n/2))]/(SampTime); %find the corresponding frequency in Hz
%%%%%%%%%%End Calculating power and frequency scale for x axis pwr is now power of standardized coefficients


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%BEGsiIN X Series Hartley test
[B,Index_dc_included]=sort(pwr,'descend'); %This sorts the array biggest to smallest and returns and Index of
where they are in the array
[B,Pos_Index_dc_included]=sort(pos_pwr,'descend'); %Pos_Index_dc_included is the positive coefficients only
[B,Index_dc_zero]=sort(pwr_dc_zero,'descend'); %This sorts the array biggest to smallest and returns and Index
of where they are in the array
[B,Index_no_dc]=sort(pwr_no_dc,'descend'); %Pos_Index is the positive coefficients only WITH DC set to ZERO
[B,Pos_Index_dc_zero]=sort(pos_pwr_dc_zero,'descend'); %Pos_Index is the positive coefficients only WITH DC set
to ZERO
[B,Pos_Index_no_dc]=sort(pos_pwr_no_dc,'descend'); %Pos_Index is the positive coefficients only WITH DC REMOVED
%PkSorted=pos_pwr(Pos_Index);%PkSorted is standardized power coeficients in order biggest to smallest pos spec
only INCLUDING DC
%PkSorted_no_dc=pos_pwr_no_dc(Pos_Index_dc_zero);%PkSorted is standardized power coeficients in order biggest
to smallest pos spec only WITHOUT DC
%why?
%%%%%%%%%%%%Calculate power in residuals for all coefficients
SigHarmonics=n; %How many significant Harmonics?
for y=1: (floor(n/2)) %modded for no dc
%PwrRes(y)=(TotPwr-PkSorted(y)); %
PwrRes_no_dc(y)=(Pos_TotPwr-pos_pwr_dc_zero(y)); %  NOT SORTED
end
%%%%%%%%%%%%End calculating power in residuals


%%%%%%%%%%%Calculate Hartley Statistic for all Coefficients
%Hartley Statistic H=Pk/(1/R)*SumResVarj compared to Fsub2,2R
R=(n-3)/2;
for y=1: (floor(n/2)) %Added over to when went to pos spectrum only WITHOUT DC
%Hart(Pos_Index(y))=PkSorted(y)/((1/R)*PwrRes(y)); %Working on only positive coeffs This is in descending ORDER
%Hart_no_dc(Pos_Index_dc_zero(y))=PkSorted_no_dc(y)/((1/R)*PwrRes_no_dc(y)); %Working on only positive coeffs
This is in descending ORDER
```

```
Hart(y)=pos_pwr_dc_zero(y)/((1/R)*PwrRes_no_dc(y));%CAN DO EITHER WAY
end

%Below just counts how many significant coeffs and calculates a percentage.

y=1;
while (y<n)
ans = Hartley(Hart(Pos_Index_dc_zero(y)),2,(n-3),alpha); %This has to go in order of the INDEX because it stops
at lastharmonic
if (ans==1)
    lastharmonic=y-1;

    y=n;%This ends the looping
    percent_significant=2*(100*((lastharmonic))/n); % multiply by 2 because lastharmonic is for one side
end
y=y+1;
end
%March 26 added sum of significant power
%Remember Pos_TotPwer already has dc power removed per above
significant_power=0;
for y=1: (lastharmonic) %Calculating total power in significant coeffs
significant_power=significant_power+pos_pwr_dc_zero(Pos_Index_dc_zero(y));
end

%if (Index_dc_zero(lastharmonic)>(floor(n/2))) %checking for even or odd I think? modded for no dc
%lastharmonic=lastharmonic-1;%with no dc this forces lastharmonic to be positive side
%end
%End Calculation of Hartley Significant Coefficients and percentage


%%%%%%%%%%%%%%%Setting all but sig coefficients equal to zero and inverting
%%%%%%%%%%%%%%%FFT them back to time Domain
%return_harmonics needs to be odd
return_harmonics=top_harmonics; %Pos only no DC, will calc other side, ifft no like DC

%%%%%%% raw_coeffs is raw fft of data  %%%%%%%%%%
raw_coeffs_filtered=raw_coeffs; %No standardize because want to reverse back to time domain praw is fft of
pressure Reminder Position 1 is DC
for t=1: n/2 %This will cycle through entire fft of pressure setting inisg POSITIVE AND NEG COEFFS to zero
            if (t>return_harmonics)%If past last sig coeff will set all values to zero
            raw_coeffs_filtered(Pos_Index_dc_zero(t))=0;% Using Index_dc_zero because this is for the 2 sided
spectrum and DC has a value
            Neg_coeff=Fmax-Pos_Index_dc_zero(t)+2;%This calcs index position of opposite freq
            raw_coeffs_filtered(Neg_coeff)=0;
            end
end

%have to filter positive with separate loop half as long

filtered_pos_pwr_dc_zero=pos_pwr_dc_zero; %Initialize before filtering
filtered_pos_pwr_no_dc=pos_pwr_no_dc;
for t=2: length(filtered_pos_pwr_dc_zero)% Runs for half spectrum plus DC
    if (t>return_harmonics)%Return_harmonics because positive only
    filtered_pos_pwr_dc_zero(Pos_Index_dc_zero(t))=0;%
    end
end
for t=1: length(filtered_pos_pwr_no_dc) % Should run 1 less because DC removed
    if (t>return_harmonics)% Minus 1 because top_harmonics wants DC
        filtered_pos_pwr_no_dc(Pos_Index_no_dc(t))=0;
    end
end


%%%raw_filtered_no_dc=filtered;              %This matches Thibos homework 9 don't know about n took it out to
keep mag
raw_coeffs_filtered_dc_zero=raw_coeffs_filtered; %
raw_coeffs_filtered_dc_zero(1)=0; %First position of praw is DC  DOING THIS TO INVERSE BACK TO TIME???
%WHY??  stdfiltered=(filtered_no_dc-mean_praw)/stdev;%This needed to make coefficients standardized Gaussian
random variable with zero
%mean and unit variance.  Squaring them then means they will then be
%distrubuted as Chi-squared
%End standardizeisng Filtered Coefficients
%pwrfiltered=abs(stdfiltered).^2;    %Need the abs to see spectrum correctly with other graphs square gives
power
filtered_pos_spectrum=filtered_pos_pwr_dc_zero(1:(floor(n/2)));%Sets equal to positive side of filtered
spectrum WITH DC removed
```

```
%Plot Power Spectrum of ALL Harmonics
figure(1) %Figure 1
semilogx(freq_no_dc,pos_pwr_no_dc,'*b');%
xlabel('Frequency (Hz)');
ylabel('Amplitude');
%v=num2str(top_harmonics);
%v='ALL';

graph_title1=['Pos side of Pwr Spectrum ALL Coefficients  '];
graph_title2=['Data File:',file_w_ext,' Alpha:',alpha_string,' Confidence Level'];
%graph_title=['Pos side of Pwr Spectrum Top  ',v,' Coefficients  ';'Data File:',file_w_ext,' Alpha:',alpha];
%twoline_title=[graph_title1.graph_title2];
title({graph_title1;graph_title2});
%title(twoline_title);
%xlim([0 freq(harmonic_order(top_harmonics))+1e11])

%Plot Power Spectrum of Top Harmonics but not Confidence Intervals
figure(2)%Figure 2
%semilogx(freq_no_dc,filtered_pos_spectrum,'*b');%Freq is off by 2 so adjusted above Shuld be using Xgrph?
semilogx(freq_no_dc,filtered_pos_pwr_no_dc,'*b');%Freq is off by 2 so adjusted above Shuld be using Xgrph?
xlabel('Frequency (Hz)');
ylabel('Amplitude');
graph_title1=['Pos side of Pwr Spectrum Top  ',v,' Coefficients  '];
graph_title2=['Data File:',file_w_ext,' Alpha:',alpha_string,' Confidence Level'];
%graph_title=['Pos Pwr Spectrum Top  ',v,' Harmonics  ',file_w_ext];
title({graph_title1;graph_title2});
%xlim([0 .4e12])


%%%%%%%Trying to plot confidence intervals (circles centered on Ahat Bhat

%Begin Confidence Interval Illustration for Hartley Sig harmonics
prob = 1 - alpha;
F2_2R = finv(prob,2,n-3);

for y=1: n/2 % added over 2 when went pos only Used to be lastharmonic but changed to see insig coeffs
   magnitude(y)=sqrt(pos_pwr_no_dc(y));
   rho(y)=sqrt((F2_2R/R)*PwrRes_no_dc(y));
    % magnitude(y)=sqrt(PkSorted_no_dc(y));
   % rho(y)=sqrt((F2_2R/R)*PwrRes_no_dc(y));
end

figure(3)  %Figure 3 This is the Graph of Circular confidence Intervals

DC=Pos_raw_coeffs_standardized(1); %This assumes DC is the first component in the array
RDC=real(DC);
IDC=imag(DC);
plot(DC,'.-y');
hold;
Coeff1=Pos_raw_coeffs_standardized(Pos_Index_dc_zero(1)); %Every other odd coefficient of whole is same as
positive side of spectrum
RCoeff1=real(Coeff1);
ICoeff1=imag(Coeff1);
plot(Coeff1,'.-c');
Coeff2=Pos_raw_coeffs_standardized(Pos_Index_dc_zero(2));
RCoeff2=real(Coeff2);
ICoeff2=imag(Coeff2);
plot(Coeff2,'.-m');
Coeff3=Pos_raw_coeffs_standardized(Pos_Index_dc_zero(3));
RCoeff3=real(Coeff3);
ICoeff3=imag(Coeff3);
plot(Coeff3,'.-g');
Coeff4=Pos_raw_coeffs_standardized(Pos_Index_dc_zero(4));
RCoeff4=real(Coeff4);
ICoeff4=imag(Coeff4);
plot(Coeff4,'.-b');

%Coeff5=praw_standardized(Index_dc_included(11));
%RCoeff5=real(Coeff5);
%ICoeff5=imag(Coeff5);
%plot(Coeff5,'.-r');

%Have to adjust for left or right side of spectrum

lh=num2str(lastharmonic);%Sets lastharmonic equal to a string that can be included in title
CLast=Pos_raw_coeffs_standardized(Pos_Index_dc_included(lastharmonic)); %Times 2 because lh is for 1/2 spectrum
RCLast=real(CLast);
```

```
ICLast=imag(CLast);
plot(CLast,'.-k');

%Below is last harmonic plus 100 to look for origin inclusion
if(single_data_set==0)
lhplus100=num2str((lastharmonic)+100);
CLastplus100=Pos_raw_coeffs_standardized(Pos_Index_dc_included(((lastharmonic)+100))); %Times 2 because lh is
for 1/2 spectrum
RCLastplus100=real(CLastplus100);
ICLastplus100=imag(CLastplus100);
plot(CLastplus100,'.-r');
legend('DC','Coeff1','Coeff2','Coeff3','Coeff4','CLast','CLastplus100');
else
lhplus1000=num2str((lastharmonic*2)+1000);
CLastplus1000=Pos_raw_coeffs_standardized(Pos_Index_dc_included(((lastharmonic*2)+1000))); %Times 2 because lh
is for 1/2 spectrum
RCLastplus1000=real(CLastplus1000);
ICLastplus1000=imag(CLastplus1000);
plot(CLastplus1000,'.-r');
legend('DC','Coeff1','Coeff2','Coeff3','Coeff4','CLast','CLastplus1000');
end
circle([RDC,IDC],rho(1),1000,'.-y');
circle([RCoeff1,ICoeff1],rho(3),1000,'.-c');
circle([RCoeff2,ICoeff2],rho(5),1000,'.-m');
circle([RCoeff3,ICoeff3],rho(7),1000,'.-g');
circle([RCoeff4,ICoeff4],rho(9),1000,'.-b');
%circle([RCoeff5,ICoeff5],rho(11),1000,'.-r');
circle([RCLast,ICLast],rho(Pos_Index_dc_included((lastharmonic))),1000,'.-k');
if (single_data_set==0)
circle([RCLastplus100,ICLastplus100],rho(Pos_Index_dc_included((lastharmonic)+100)),1000,'.-r');
else
   lhp1000=(lastharmonic)+1000;
   if mod(((lastharmonic)+1000),2)
   %disp('odd')
   lhp1000=(lastharmonic)+1000;
   else
   %disp('even')
   lhp1000=(lastharmonic)+999;

   end
if (Index_dc_included(lhp1000) > length(rho))
    disp (lhp1000)

   lhp1000=lhp1000-1;
end
circle([RCLastplus1000,ICLastplus1000],rho(Pos_Index_dc_included(lhp1000)),1000,'.-r');
end
axis square;
axis equal;
grid on;
%v=num2str(top_harmonics);

graph_title1=['Geometric Illustraion of Confidence Intervals for Top Four Coefficients'];
graph_title2=['For Data File:',file_w_ext,' Alpha: ',alpha_string];
title({graph_title1;graph_title2});
xlabel('Real Part of Measured Coefficient (X Axis)') ;
ylabel('Imaginary Part of Measured Coefficient (Y Axis)');
if (single_data_set==0)
legend('DC','Coeff1','Coeff2','Coeff3','Coeff4',lh,lhplus100);
else
legend('DC','Coeff1','Coeff2','Coeff3','Coeff4',lh,lhplus1000);
end

figure(4) % Time Domain of Top Harmonics
% Want to include pos and neg freq of top harmonics without DC
%Filter out all but top harmonics
%TimeSignal_harmonics(1)=0;
%NEW METHOD
y=1;
%TimeSignal_harmonics=raw_coeffs;
TimeSignal_harmonics(length(raw_coeffs))=0;
while (y<(floor(n/2)))
   if (y<return_harmonics+1)
    TimeSignal_harmonics(Pos_Index_dc_zero(y))=raw_coeffs(y);
     Neg_coeff=(Fmax-Pos_Index_dc_zero(y)+2);%This calcs index position of opposite freq
    TimeSignal_harmonics(Neg_coeff)=conj(raw_coeffs(y)); %This sets opposite freq equal
    Pos_Index_dc_zero(y)
```

```
    Neg_coeff
   end
    y=y+1;
    %TimeSignal_harmonics(Pos_Index_dc_zero(y))=0;
    %Neg_coeff=((Pos_Index_dc_zero(y))+1)+((Fmax-1)/2);%This calcs index position of opposite freq
    %TimeSignal_harmonics(Neg_coeff)=0; %This sets opposite freq equal
end

%for y=1:return_harmonics-1
%TimeSignal_harmonics(Index_no_dc(y))=raw_coeffs(Index_no_dc(y));
%end

%TimeSignal=ifft(filtered); %Filtered is raw with small coeffs zeroed out
%raw_coeffs_filtered(1)=0;
TimeSignal=ifft(TimeSignal_harmonics);
plot(TimeSignal);
%v=num2str(top_harmonics);
graph_title1=['Time Domain of Top  ',v,' Harmonic Pressures '];
graph_title2=['For Data File: ',file_w_ext,' Alpha: ',alpha_string];
title({graph_title1;graph_title2});
ylabel('Press Amplitude');
xlabel('Real Time periodic Pressure');

%%%%%%Begin Hartley Summation Report
outfile=[file_name '_stats.txt'];
report_1=fopen(outfile,'w');

%Below prints sig coeffs to csv file for matching
outfile2=[file_name '_data.txt'];
report_2=fopen(outfile2,'w');
for p=1:1:lastharmonic %used to be 59 and 61 added 20 to each
fprintf(report_2,'%2.4E %s %4.4f %s
%4.4f\n',freq_no_dc(Pos_Index_no_dc(p)),',',Hart(Pos_Index_dc_zero(p)),',',pos_pwr_dc_zero(Pos_Index_dc_zero(p)
));
end
fclose(report_2);
%End printing out csv file

disp(' ');
fprintf(report_1,'-------------------------------------------\n');
fprintf(report_1,'Analysis for data file:    ');
fprintf(report_1,file_w_ext);
fprintf(report_1,'\n');
fprintf(report_1,'-------------------------------------------\n');
fprintf(report_1,'Total Number of Coefficients\n');
fprintf(report_1,'%4.1i \n\n',n);
fprintf(report_1,'Total Number of Significant Coefficients\n');
fprintf(report_1,'%4.1i \n\n',lastharmonic-1);
fprintf(report_1,'Percentage of Total that are Significant\n');
fprintf(report_1,'%3.5f \n\n',percent_significant);
fprintf(report_1,'Power in Significant Coefficients div Total Power*100\n');
fprintf(report_1,'%3.5f \n\n',100*significant_power/Pos_TotPwr);
fprintf(report_1,'Upper and Lower Frequency Limits Returned by Transform\n');
fprintf(report_1,'Lower Frequency      Upper Frequency     Sampling Frequency\n');
fprintf(report_1,'%2.3E              %2.3E        %2.5E\n\n',fbottom,ftop,Fsamp);
fprintf(report_1,'F2,2R                          Alpha\n');
fprintf(report_1,'%4.2f                           %2.2G\n\n',F2_2R,alpha);
fprintf(report_1,'Top 100 Coefficients Ranked by Magnitude \n\n');
fprintf(report_1,'  Hz        Hartley    Power       Hz        Hartley    Power\n\n');
%sorted=1;
%for fcnt=1:length(pwr_dc_zero)
%if (Index_dc_included(fcnt)<floor(n/2))
%sorted_freq(sorted)=freq_no_dc(Index_dc_included(fcnt));
%sorted_pwr(sorted)=pwr_dc_zero(Index_dc_included(fcnt));
%sorted_hart(sorted)=Hart(Index_dc_included(fcnt));
%sorted=sorted+1;
%end
%end
na='  n/a';
fprintf(report_1,'%2.3E   %s    %4.4f\n',freq(1),na,pos_pwr(1));
for p=1:1:50 %used to be 59 and 61 added 20 to each
fprintf(report_1,'%2.3E    %4.3f    %4.4f    %2.3E    %4.3f
%4.4f\n',freq_no_dc(Pos_Index_no_dc(p)),Hart(Pos_Index_dc_zero(p)),pos_pwr_dc_zero(Pos_Index_dc_zero(p)),freq_n
o_dc(Pos_Index_no_dc(p+50)),Hart(Pos_Index_dc_zero(p+50)),pos_pwr_dc_zero(Pos_Index_no_dc(p+50)));
%Want to save data to file appropriate for spreadsheet analysis
%summary[p,p,p]=[freq(Pos_Index_dc_included(p)),Hart(Pos_Index(p)),pos_pwr(Pos_Index(p))];
%summary[p+50,p+50,p+50]=[freq(Pos_Index(p+50)),Hart(Pos_Index(p+50)),pos_pwr(Pos_Index(p+50))];
```

```
%Cant get brackets to work?!!
end
fclose(report_1);
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%END Hartley Test and report%%%%%%%%%%%



%if (single_data_set==0)%Only do confidence bounds if data is decimated or Windowed
if (single_data_set==99)%Throwing out this graph no meaning anyway
%Begin finding confidence bounds by taking mean and stdev of dec columns
%Need to start by determining top 5 coeffs position
row_length=length(file_data(1,:));%file_data is (16 by 6249)
for t=1:sample_size  % 1 to 16
row_fft(t,:)=fft(file_data(t,:)); %Get coeffs of each column resulting in sampl_size vectors
%The mean of row_fft(:,x) equals praw(x) Don't need row_fft
end
%%%%%%%%%%%%%%%%%%%%%%%%%  Calc Std Err Mean for conf intervals
mean_coefficients=mean(row_fft); %This equals praw which is fft of the avg data GOOD CROSSCHECK TEST
%Standardize Coefficeints by subtracting off mean and dividing by Standard
%deviation USE SAME mean and sdev as orig calc to keep centered
% from above stdev is standard deviation of praw
mean_coefficients_standardized=(row_fft-mean_raw_coeffs)/stdev;%This needed to make coefficients standardized
Gaussian random variable with zero mean
%consistent with first calculations of power for the first graphs
pwr_coeffs=abs(mean_coefficients_standardized).^2; %Changes standardized fft coefs into powers
%NOW can get std error of the mean
%First need to get pos only no dc

%pos_pwr_no_dc=pwr(2:(floor(n/2)+1));%%%%Positive half of spectrum only with f1 in first position

pos_pwr_coeffs=pwr_coeffs(2:(floor(n/2)+1)); %Positive half with f1 in first position
for rank=1:top_harmonics+1
coefficient_stdev(rank)=std(pos_pwr_coeffs(:,Pos_Index_no_dc(rank)));
coefficient_average(rank)=mean(pos_pwr_coeffs(:,Pos_Index_no_dc(rank))); %crosscheck should be close to
filtered_pos_pwr_no_dc
%SEM is s/sqrt(n) where s is sample stdev and n is sample size
coef_serr_mean(rank)=coefficient_stdev(rank)/sqrt(sample_size);
end
%%%%%%%%%%%%%%%%Confidence Intervals%%%%%%%%%%%%%%%
%Want to know 99.999%confidence interval on the sample mean?
%We need the sample mean, standard dev s, sample size n and df n-1
%Then plugin to MATLAB tinv(probability,degrees of freedom)
%t-value for alpha/2 and 16-1=15df is tinv(0.9999,15)
%Limits are t-value*.01/sqrt(N)
for rank=1:top_harmonics+1
t_value(rank)=tinv(1-alpha,sample_size-1);
conf_bound(rank)=t_value(rank)*pos_pwr_coeffs(rank)/sqrt(16);
end

%figure(2)%Figure 2
%semilogx(freq,filtered_pos_spectrum,'*b');%Freq is off by 2 so adjusted above
figure(2) %Bring back figure 2 to add confidence bounds
hold;
adjust=0;%Had a problem early on prolly don't need anymore

harm=1;
semilogx(freq_no_dc(Pos_Index_no_dc(harm)-adjust),abs(filtered_pos_pwr_no_dc(Pos_Index_no_dc(harm)-adjust))-
conf_bound(harm),'*r');%Freq is off by 2 so adjusted above
semilogx(freq_no_dc(Pos_Index_no_dc(harm)-adjust),abs(filtered_pos_pwr_no_dc(Pos_Index_no_dc(harm)-
adjust))+conf_bound(harm),'*r');%Freq is off by 2 so adjusted above
harm=2;
semilogx(freq_no_dc(Pos_Index_no_dc(harm)-adjust),abs(filtered_pos_pwr_no_dc(Pos_Index_no_dc(harm)-adjust))-
conf_bound(harm),'*r');%Freq is off by 2 so adjusted above
semilogx(freq_no_dc(Pos_Index_no_dc(harm)-adjust),abs(filtered_pos_pwr_no_dc(Pos_Index_no_dc(harm)-
adjust))+conf_bound(harm),'*r');%Freq is off by 2 so adjusted above
harm=3;
semilogx(freq_no_dc(Pos_Index_no_dc(harm)-adjust),abs(filtered_pos_pwr_no_dc(Pos_Index_no_dc(harm)-adjust))-
conf_bound(harm),'*r');%Freq is off by 2 so adjusted above
semilogx(freq_no_dc(Pos_Index_no_dc(harm)-adjust),abs(filtered_pos_pwr_no_dc(Pos_Index_no_dc(harm)-
adjust))+conf_bound(harm),'*r');%Freq is off by 2 so adjusted above
harm=4;
semilogx(freq_no_dc(Pos_Index_no_dc(harm)-adjust),abs(filtered_pos_pwr_no_dc(Pos_Index_no_dc(harm)-adjust))-
conf_bound(harm),'*r');%Freq is off by 2 so adjusted above
semilogx(freq_no_dc(Pos_Index_no_dc(harm)-adjust),abs(filtered_pos_pwr_no_dc(Pos_Index_no_dc(harm)-
adjust))+conf_bound(harm),'*r');%Freq is off by 2 so adjusted above
harm=5;
```

```
semilogx(freq_no_dc(Pos_Index_no_dc(harm)-adjust),abs(filtered_pos_pwr_no_dc(Pos_Index_no_dc(harm)-adjust))-
conf_bound(harm),'*r');%Freq is off by 2 so adjusted above
semilogx(freq_no_dc(Pos_Index_no_dc(harm)-adjust),abs(filtered_pos_pwr_no_dc(Pos_Index_no_dc(harm)-
adjust))+conf_bound(harm),'*r');%Freq is off by 2 so adjusted above
%harm=6;
%semilogx(freq_no_dc(Pos_Index_no_dc(harm)-adjust),abs(filtered_pos_pwr_no_dc(Pos_Index_no_dc(harm)-adjust))-
2*coef_serr_mean(harm),'*r');%Freq is off by 2 so adjusted above
%semilogx(freq_no_dc(Pos_Index_no_dc(harm)-adjust),abs(filtered_pos_pwr_no_dc(Pos_Index_no_dc(harm)-
adjust))+2*coef_serr_mean(harm),'*r');%Freq is off by 2 so adjusted above


%xlim([1e10 1e12])

%End finding confidence bounds

end %This closes single_data_set test for confidence bounds
```

circle.m

(This program is a small subroutine called from within the Hartley_Combo_Final.m program to draw the circles on the confidence interval graphs.)

```
function H=circle(center,radius,NOP,style)
%-------------------------------------------------------------------------------------------------
% H=CIRCLE(CENTER,RADIUS,NOP,STYLE)
% This routine draws a circle with center defined as
% a vector CENTER, radius as a scaler RADIS. NOP is
% the number of points on the circle. As to STYLE,
% use it the same way as you use the rountine PLOT.
% Since the handle of the object is returned, you
% use routine SET to get the best result.
%
%   Usage Examples,
%
%   circle([1,3],3,1000,':');
%   circle([2,4],2,1000,'--');
%
%   Zhenhai Wang <zhenhai@ieee.org>
%   Version 1.00
%   December, 2002
%-------------------------------------------------------------------------------------------------

if (nargin <3),
 error('Please see help for INPUT DATA.');
elseif (nargin==3)
    style='b-';
end;
THETA=linspace(0,2*pi,NOP);
RHO=ones(1,NOP)*radius;
[X,Y] = pol2cart(THETA,RHO);
X=X+center(1);
Y=Y+center(2);
H=plot(X,Y,style);
%axis square;
```

Install_test_signal_to_Orig_Data_xyz.m

(This program was written to verify proper scaling of the X-axis on all power spectrum graphs produced by the analysis. It was used to install frequencies of several recognizable values and varying magnitudes into actual system pressures. Those frequencies could then be located in the power spectrum and used to verify the chain of data input, calculations and data output.)

```
clear all
%%This is where the actual data is normally read in
load ../runs/prod/run_combined/x_sumation_pressures.txt
load ../runs/prod/run_combined/y_sumation_pressures.txt
load ../runs/prod/run_combined/z_sumation_pressures.txt
pressurex=x_sumation_pressures(:,3);
pressurey=y_sumation_pressures(:,3);
pressurez=z_sumation_pressures(:,3);
n = length(pressurex);
%ny = length(pressurey);
%nz = length(pressurez);
%runtime_partial=x_runtime_pressures(:,2);
%ewald_partial=x_ewald_pressures(:,2);
%times=xdata_50(:,1);

%This is where synthetic data is made instead of actual data
%The synth data must be based on only 2 variables sampling interval and
%number of samples just like the simulation results
%%%%%%%%%Begin Signal Install
f1 = 3.3333e9;  %Frequency in Hz of test signal 1 (1tera hz)
f2 = 3.3333e10;
f3 = 3.3333e11;
f4 = 3.3333e12;  %Frequency in Hz of test signal 1 (1tera hz)
f5 = 3.3333e13;
f6 = 3.3333e14;
k1=0;
k2=0;
k3=5e2;
k4=0;
k5=0;
k6=0;

spr=10; %pressureprofile output was 10 steps per cycle during simulation
Tint = 2e-15*spr;  %sampling time interval in s 1E-3=1ms or 2fs ORIGINAL
%ftop = (1/2)*1/Tint; %simple nyquist frequency
%fbottom = 1/(n*Tint); %2 over the window
%SampTimeActual=n*Tint;
SampTime = (Tint*n); %Total time interval
t= 0:Tint:(((n*Tint)-Tint)); % Setup a time vector
Fsamp=1/Tint; %sampling frequency is 1/interval samples every second in Hz or sampling rate
%s=square(t);
signal =
(0+k1*sin(2*pi*f1*t)+k2*sin(2*pi*f2*t)+k3*sin(2*pi*f3*t)+k4*sin(2*pi*f4*t)+k5*sin(2*pi*f5*t)+k6*sin(2*pi*f6*t))
';


%%%%%%%%%End Signal Install
%%%%%f1 = 5e10;  %Frequency in Hz of test signal 1 (1tera hz)
%%%%%f2 = 1e12;
%%%%%f3 = 1e13;

x1=pressurex;
y1=pressurey;
z1=pressurez;
%pressurex(length(signal))=0;
%pressurey(length(signal))=0;
%pressurez(length(signal))=0;
%y1=x1;  %This retains the original values
```

```
pressurex=x1+signal;  %
pressurey=y1+signal;  %
pressurez=z1+signal;  %
%pressurex=signal;
x_sumation_pressures(:,3)=pressurex;
y_sumation_pressures(:,3)=pressurey;
z_sumation_pressures(:,3)=pressurez;
%save Linearcolumn16decx_sig.txt -ascii column16

save ../runs/prod/run_combined/x_sumation_pressures_sig_e11.txt -ascii x_sumation_pressures
save ../runs/prod/run_combined/y_sumation_pressures_sig_e11.txt -ascii y_sumation_pressures
save ../runs/prod/run_combined/z_sumation_pressures_sig_e11.txt -ascii z_sumation_pressures
```

APPENDIX-G: MODEL CONSTRUCTION PROCEDURE

Construct ab-initio molecular model

1. Take the chosen base pair sequence and construct an ab-initio molecular model. This is done with code found in the AMBER suite of tools called NAB or "Nucleic Acid Builder". A web-based implementation is available at http://structure.usc.edu/make-na/server.html. (Stroud, 2006) This server was intended primarily for crystallographers but works well for the simple model needed for this investigation. The sequence TATAAACGCC is input as the A chain TOP segment and the reverse is input as the B chain BOTTOM segment. Helix type B is selected and both chain A and segment B are set to type DNA. In the advanced options section Asterisks' are set to represent sugar atoms and hydrogen's are set to not be included in the model. Hydrogens will be added later with the structure file generator within VMD. Chain IDs A and B are set and a PDB file type is returned. This molecule file was labeled 1TATAAACGCC_raw.pdb.

2. The entire procedure is repeated exactly except the chain IDs are changed to C and D to allow combination of the 2 models into one with the VMD modeling program. The second model was generated and saved as 2TATAAACGCC_raw.pdb.

3. Two files are needed to run MD simulations with NAMD. The atomic coordinate file (.pdb) and the structure file (.psf) containing bonding interaction information. We will use the VMD autopsf generator feature to create an appropriate structure file. First we run VMD and load the 1TATAAACGCC_raw.pdb file. From the Extensions menu select Modeling>Automatic PSF Builder.

4. Change the output basename to 1TATAAACGCC. Click Load Input Files. The autopsf builder defaults to the top_all27_protein_lipid_na.inp topology file which works fine for a simple nucleic acid helix.

5. Select EVERYTHING to be included in the PSF and PDB files.

6. Click guess and split chains using current selections.

7. Click Create chains. This causes autopsf to write out the 2 identified chains into 2 seperate temporary pdb files for combination in the last step. Rename N1 and N2 to N3 and N4 for the second molecule.

8. Click apply patches and finish PSF/PDB to complete generation of the .psf and .pdb input files. The resulting molecule looks like this:

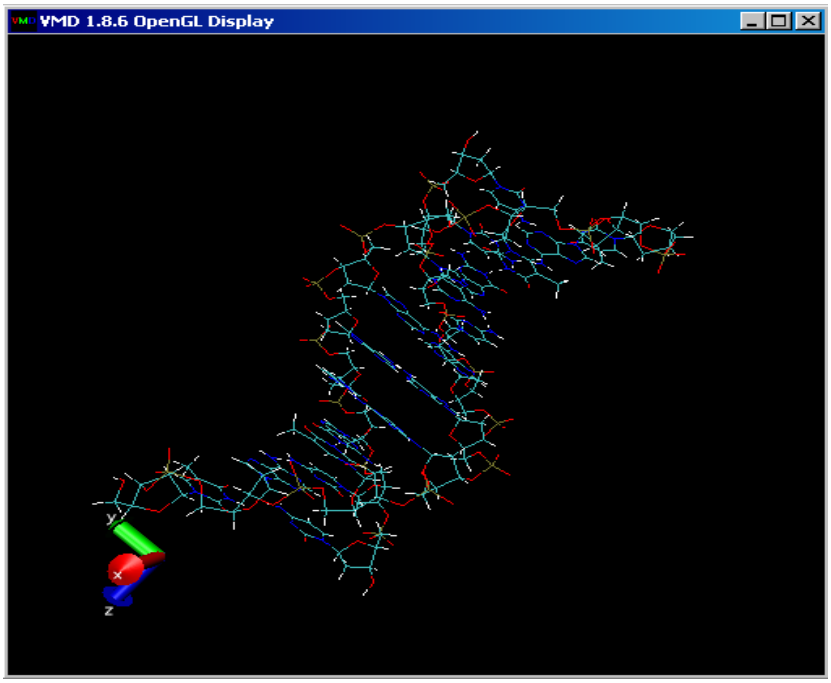


Figure 55: Sample PSF gen

9. Close VMD to clear memory and repeat the process for the 2TATAAACGCC_raw.pdb file as well. Rename N1 and N2 chains to N3 and N4. We now have 2 separate models of the same molecule with different chain names. We can now combine them into the geometric configurations needed for analysis.

Combining 2 DNA models into a single system

1. We will use VMD's Tkc console to accomplish this for each configuration. We will start with 10mer_linear_0.pdb. Begin by creating a tcl script with the following commands:

```
set psf0 ./1tataaacgcc.psf
set pdb0 ./1tataaacgcc.pdb
set psf1 ./2tataaacgcc.psf
set pdb1 ./2tataaacgcc.pdb
set finalPsf 10mer_linear_0_double.psf
set finalPdb 10mer_linear_0_double.pdb
package require psfgen
resetpsf
readpsf $psf0
coordpdb $pdb0
readpsf $psf1
coordpdb $pdb1
writepdb $finalPdb
writepsf $finalPsf
```

Move Molecule 2 forty five angstroms in the +z direction to achieve 10 angstrom space

1. Open TKconsole
2. Enter set sel [atomselect top "segname CH3"]
3. Enter (selection name returned) atomselect1 moveby (0 0 45)
4. set sel [atomselect top "segname CH4"]

5. Enter (selection name returned) atomselect2 moveby

6. Enter set all [atomselect top all] (selection name atomselectX will be returned)

7. Enter atomselectX writepdb 10mer_linear_0.pdb

8. Enter atomselectX writepsf 10mer_linear_0.psf

9. Repeat this for each geometric permutation. The result looks like this.

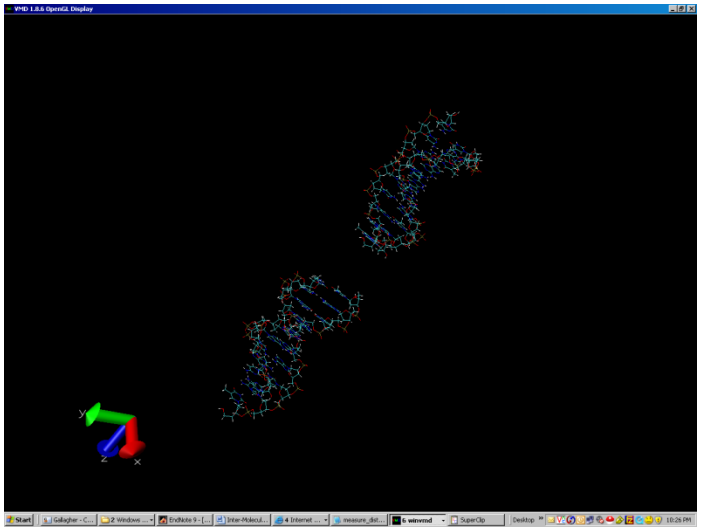


Figure 56: Combining Procedure Result

Solvate the System

1. Create a script file called solvatesystem.tcl and input the following lines into a txt file and save to a working directory.

package require solvate

solvate ../test_sequence/10mer_linear_0.psf ../test_sequence/10mer_linear_0.pdb +z 7 -z 7 +x 12 -x 12 +y 12 -y 12 -o 10mer_linear_0_water

2. In the TK console Enter source ../solvatesystem.tcl to get the following;

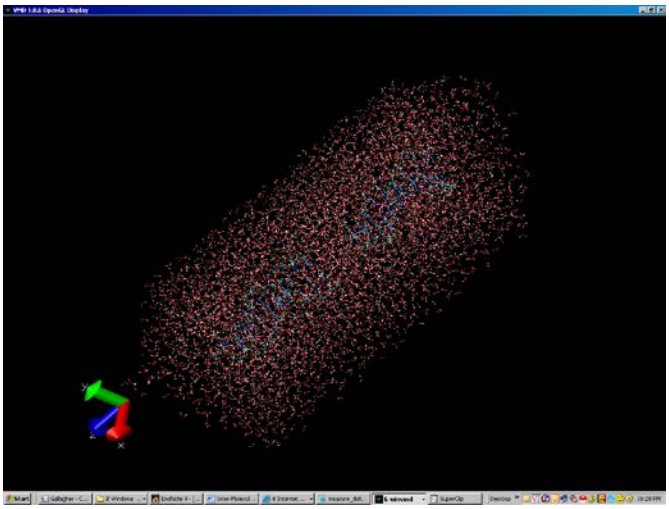


Figure 57: Solvation Results

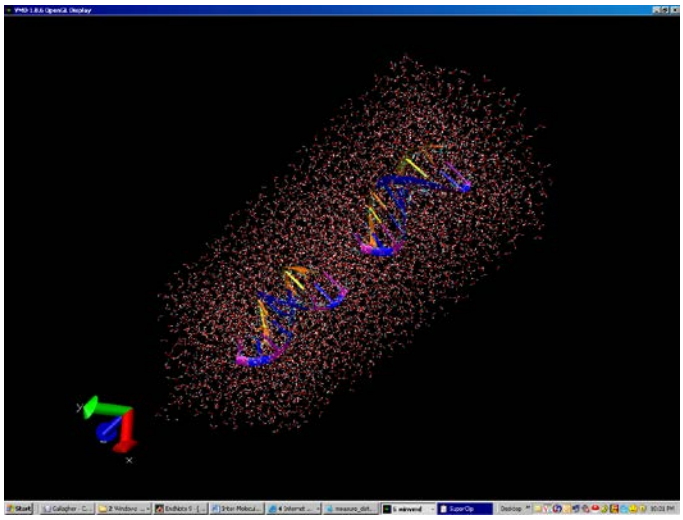


Figure 58: Solvated Molecule Better View

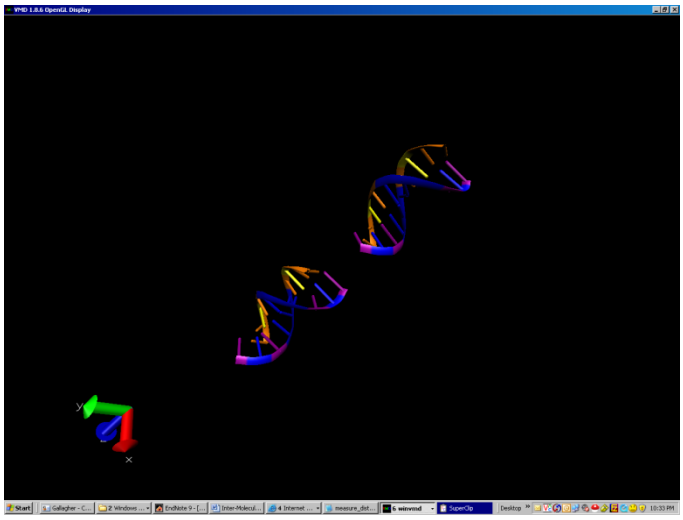


Figure 59: Solvated Molecule Ribbon View no Water

Ionize the System

3. Next Ionize the system by clearing VMD and reloading.

4. Select Extensions>TK Console from the VMD main screen.

5. At the TK Console prompt change to the directory where the 2 molecule files are located. In this example its cd Test_sequence.

6. Were going to Ionize the system by adding Na and Cl atoms until the net charge in the system is zero and the average ionic concentration of the system is 0.5 mol/L, any less and the charge would be too low to add Cl ions. This is necessary because the simulation is going to use particle-mesh Ewald (PME) summation which requires the system to be electrically neutral.

7. To perform the ionization type in autoionize -psf 10mer_linear_0_water.psf -pdb 10mer_linear_0_water.pdb -is 0.5 -o ionized –from(min dist from mole) 5.0 –between(min dist between ions) 5.0 The results are:

>Main< (TEST_Sequence) 52 % autoionize -psf 10mer_linear_0_water.psf -pdb 10mer_linear_0_water.pdb -is 0.5 -o ion

reading topology file C:/Program Files/University of

Illinois/VMD/plugins/noarch/tcl/autoionize1.2/ions.top

>>>>>> SOD and CLA Ions Topology File <<<<<<<<<

extracted from

>>>>>>>>CHARMM22 All-Hydrogen Topology File for Proteins <<<<<<

>>>>>>>>>>>>>>>>>>>> August 1999 <<<<<<<<<<<<<<<<<<<<<<<<<<<<

>>>>>>> Direct comments to Alexander D. MacKerell Jr. <<<<<<<<<

>>>>>> 410-706-7442 or email: alex,mmiris.ab.umd.edu <<<<<<<<<

Created by CHARMM version 27 1

Autoionize) Reading 10mer_linear_0_water.psf/10mer_linear_0_water.pdb...

clearing structure, preserving topology and aliases

reading structure from psf file 10mer_linear_0_water.psf

psf file does not contain cross-terms

reading coordinates from pdb file 10mer_linear_0_water.pdb

Autoionize) System net charge before adding ions: -35.9999972709e

Autoionize) Desired ion concentration 0.5 mol/L

Autoionize) Adding 42 Na and 6 Cl ions, total 48 ions

Autoionize) Required min distance from molecule 5A

Autoionize) Required min distance between ions 5A

Autoionize) Output file prefix 'ion'

Autoionize) Obtained positions for 48 ions

Autoionize) Tagged 48 water molecules for deleting

Autoionize) Deleted 48 water molecules

Autoionize) Adding 42 SOD and 6 CLA residues...

building segment ION

setting patch for first residue to NONE

setting patch for last residue to NONE

Info: generating structure...

Info: segment complete.

Autoionize) Randomizing ion positions...

Autoionize) Assigned 42 Na coordinates

Autoionize) Assigned 6 Cl coordinates

Info: writing psf file ion.psf

total of 16757 atoms

total of 11654 bonds

total of 5151 angles

total of 0 dihedrals

total of 0 impropers

total of 0 cross-terms

Info: psf file complete.

Info: writing pdb file ion.pdb

Info: pdb file complete.

Autoionize) Reloading the system with added ions...

Autoionize) System net charge after adding ions: 2.72919560729e-006e

Autoionize) All done.

>Main< (TEST_Sequence) 53 %

8. The program assigns 42 Na coordinate positions and 6 Cl coordinate positions adding a total of 48 ions randomly and generating the necessary files.

9. After adjusting the graphical representation settings and display type settings per the screens below the results looks like this.

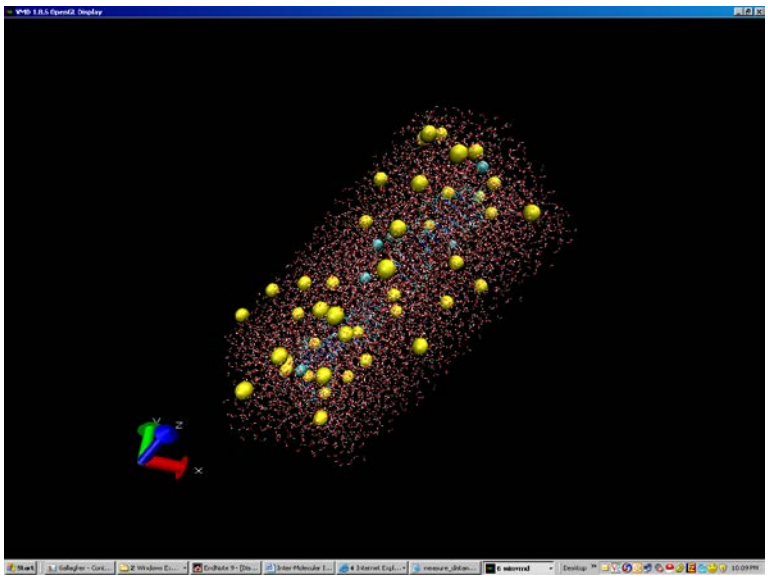


Figure 60: Ions Highlighted in Yellow

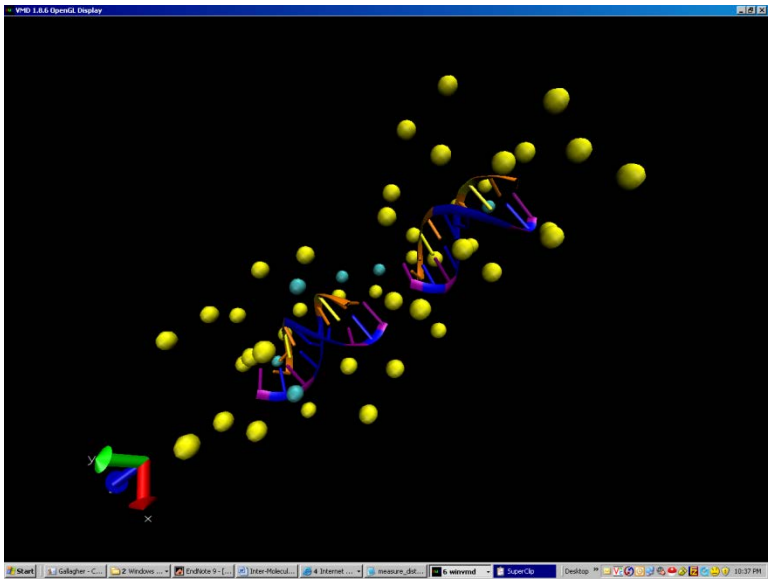


Figure 61: Ions and Molecule Better View no Water

10. Now we need to calculate the center of the system as well as the minimum and maximum coordinates in the X, Y, and Z directions. Start by entering the following commands into the TK Console;

>Main< (10mer_linear_0) 56 % set everyone [atomselect top all]

atomselect0

>Main< (10mer_linear_0) 57 % measure center $everyone

0.0332483612001 -0.0598412193358 37.7706336975

>Main< (10mer_linear_0) 58 % measure minmax $everyone

(-21.8020000458 -22.2549991608 -10.2740001678) (21.7999992371 22.2859992981 85.7229995728)

>Main< (10mer_linear_0) 59 %

From the results we know the cell origin is 0A , -0.1A, 37.7A and the cell basis vectors should be 38.6x 39.6y 91.0z (5A less then total edge length to avoid vacuum).

Generation of Phix174 molecular systems:

chain ID AB

176-469 chain ID CD

406-699 chain ID EF

406-699 chain ID GH

Loaded into one system configured as

AB CD

EF GH

Translated AB(n1 n2) 0 50 0

Translated CD(n3 n4) 50 50 0

Translated EF(n5 n6) 0 0 0

Translated GH (n6 n7) 50 0 0

1. Open TKconsole
2. Enter set sel [atomselect top “segname N1”]
3. Enter (selection name returned) atomselect1 moveby {0 50 0}
4. set sel [atomselect top “segname N2”]
5. Enter (selection name returned) atomselect2 moveby {0 50 0}
6. Enter set all [atomselect top all] (selection name atomselectX will be returned)
7. Enter atomselectX writepdb phix174_176_469_AB.pdb
8. Enter atomselectX writepsf phix174_176_469_AB.psf

9. Repeated this 4 times.

NOTE: JUST USE PDB FILE TO MANIPULATE and TRANSLATE Molecules, the display DOES NOT UPDATE, just ignore the display the molecules are moved. Then run autopsf after all the molecules are where they are supposed to be otherwise NAMD will crash with error “no VDW params for molecule type C4’.

Then used this script to combine 4 molecules. Called combine_4_mols.tcl

```
set psf0 ./phix174_176_469_AB_0.50.0.psf
set pdb0 ./phix174_176_469_AB_0.50.0.pdb
set psf1 ./phix174_176_469_CD_50.50.0.psf
set pdb1 ./phix174_176_469_CD_50.50.0.pdb
set psf2 ./phix174_406_699_EF_0.0.0.psf
set pdb2 ./phix174_406_699_EF_0.0.0.pdb
set psf3 ./phix174_406_699_GH_50.0.0.psf
set pdb3 ./phix174_406_699_GH_50.0.0.pdb
set finalPsf phix174_final_s_i.psf
set finalPdb phix174_final_s_i.pdb

package require psfgen
resetpsf

readpsf $psf0
coordpdb $pdb0
readpsf $psf1
coordpdb $pdb1
readpsf $psf2
coordpdb $pdb2
readpsf $psf3
coordpdb $pdb3

writepdb $finalPdb
writepsf $finalPsf

Created a script called autoionize_.5nacl5a.tcl and it came out too low with
no CL ions.  Have to redo.

Final file is now called ionized.pdb and ionized.psf
>Main< (50A_Seperation) 56 % source ionizesystem_.6mlnacl.tcl
reading topology file C:/Program Files/University of
Illinois/VMD/plugins/noarch/tcl/autoionize1.2/ions.top

>>>>>>       SOD and CLA Ions Topology File           <<<<<<<<<
extracted from
>>>>>>>>CHARMM22 All-Hydrogen Topology File for Proteins <<<<<<
>>>>>>>>>>>>>>>>>>>> August 1999 <<<<<<<<<<<<<<<<<<<<<<<<<<<<
>>>>>>> Direct comments to Alexander D. MacKerell Jr. <<<<<<<<<
```

```
>>>>>> 410-706-7442 or email: alex,mmiris.ab.umd.edu  <<<<<<<<<

Created by CHARMM version 27 1

Autoionize) Reading phix174_final_i.psf/phix174_final_i.pdb...
clearing structure, preserving topology and aliases
reading structure from psf file phix174_final_i.psf
psf file does not contain cross-terms
reading coordinates from pdb file phix174_final_i.pdb

Autoionize) System net charge before adding ions: -2343.999827772379e
Autoionize) Desired ion concentration 0.6 mol/L
Autoionize) WARNING: ion concentration too low, cannot add Cl ions!
Autoionize) Adding 2343 Na and 0 Cl ions, total 2343 ions
Autoionize) Required min distance from molecule 5.0A
Autoionize) Required min distance between ions 5.0A
Autoionize) Output file prefix 'ionized'
Autoionize) Obtained positions for 2343 ions
Autoionize) Tagged 2343 water molecules for deleting
Autoionize) Deleted 2343 water molecules
Autoionize) Adding 2343 SOD and 0 CLA residues...
building segment ION
setting patch for first residue to NONE
setting patch for last residue to NONE
Info: generating structure...
Info: segment complete.
Autoionize) Randomizing ion positions...
Autoionize) Assigned 2343 Na coordinates
Autoionize) Assigned 0 Cl coordinates
Info: writing psf file ionized.psf
total of 583964 atoms
total of 418530 bonds
total of 315519 angles
total of 212936 dihedrals
total of 6404 impropers
total of 0 cross-terms
Info: psf file complete.
Info: writing pdb file ionized.pdb
Info: pdb file complete.
Autoionize) Reloading the system with added ions...

Autoionize) System net charge after adding ions: -0.9998277723789215e
Autoionize) All done.
>Main< (50A_Seperation) 57 %
```

APPENDIX-H: EXCERPT OF LINEAR PDB (PROTEIN DATA BASE) FILE

(This file contains 16386 lines, one for each atom)

```
CRYST1   33.608   34.584  164.846  90.00  90.00  90.00 P 1           1
ATOM      1  C4' THY N   1       2.695  -7.020  -2.053  1.00  0.00      N1   C
ATOM      2  H4' THY N   1       3.531  -7.134  -2.590  1.00  0.00      N1   H
ATOM      3  O4' THY N   1       2.477  -5.630  -1.823  1.00  0.00      N1   O
ATOM      4  C1' THY N   1       2.308  -5.333  -0.452  1.00  0.00      N1   C
ATOM      5  H1' THY N   1       3.153  -4.901  -0.136  1.00  0.00      N1   H
ATOM      6  C2' THY N   1       2.148  -6.679   0.247  1.00  0.00      N1   C
ATOM      7  H2' THY N   1       2.595  -6.635   1.140  1.00  0.00      N1   H
ATOM      8 H2'' THY N   1       1.185  -6.946   0.218  1.00  0.00      N1   H
ATOM      9  H5T THY N   1      -0.390  -8.147  -2.267  1.00  0.00      N1   H
ATOM     10  O5' THY N   1       0.427  -7.826  -1.788  1.00  0.00      N1   O
ATOM     11  C5' THY N   1       1.443  -7.510  -2.756  1.00  0.00      N1   C
ATOM     12  H5' THY N   1       1.134  -6.739  -3.313  1.00  0.00      N1   H
ATOM     13 H5'' THY N   1       1.716  -8.349  -3.227  1.00  0.00      N1   H
ATOM     14  N1  THY N   1       1.098  -4.468  -0.370  1.00  0.00      N1   N
ATOM     15  C6  THY N   1      -0.158  -5.012  -0.400  1.00  0.00      N1   C
ATOM     16  H6  THY N   1      -0.438  -5.969  -0.473  1.00  0.00      N1   H
ATOM     17  C2  THY N   1       1.304  -3.113  -0.265  1.00  0.00      N1   C
ATOM     18  O2  THY N   1       2.414  -2.612  -0.237  1.00  0.00      N1   O
ATOM     19  N3  THY N   1       0.159  -2.344  -0.191  1.00  0.00      N1   N
ATOM     20  H3  THY N   1       0.295  -1.266  -0.106  1.00  0.00      N1   H
.

.

.

ATOM  16368  SOD SOD I  70       4.151 -15.315  99.529  1.00  0.00      ION NA
ATOM  16369  SOD SOD I  71      -8.856  10.712  92.551  1.00  0.00      ION NA
ATOM  16370  SOD SOD I  72       6.703  12.241 104.488  1.00  0.00      ION NA
ATOM  16371  SOD SOD I  73      12.153   4.430 126.029  1.00  0.00      ION NA
ATOM  16372  SOD SOD I  74       9.677  10.261  -1.284  1.00  0.00      ION NA
ATOM  16373  SOD SOD I  75       7.889  16.245  11.931  1.00  0.00      ION NA
ATOM  16374  SOD SOD I  76       8.884  -5.833  96.236  1.00  0.00      ION NA
ATOM  16375  SOD SOD I  77      -2.728  -9.966 120.694  1.00  0.00      ION NA
ATOM  16376  SOD SOD I  78       3.380  12.141 109.358  1.00  0.00      ION NA
ATOM  16377  SOD SOD I  79      14.295   9.898  46.652  1.00  0.00      ION NA
ATOM  16378  SOD SOD I  80      -5.875  -2.542  -4.595  1.00  0.00      ION NA
ATOM  16379  SOD SOD I  81       6.873 -15.850  53.679  1.00  0.00      ION NA
ATOM  16380  SOD SOD I  82      -1.050  16.353 131.671  1.00  0.00      ION NA
ATOM  16381  CLA CLA I  83      13.894  -1.879  39.128  1.00  0.00      ION CL
ATOM  16382  CLA CLA I  84     -11.625   3.598 147.913  1.00  0.00      ION CL
ATOM  16383  CLA CLA I  85      11.257  -7.051  40.960  1.00  0.00      ION CL
ATOM  16384  CLA CLA I  86      -2.924  15.121  -1.260  1.00  0.00      ION CL
ATOM  16385  CLA CLA I  87      -0.115   8.937  63.396  1.00  0.00      ION CL
ATOM  16386  CLA CLA I  88      -6.767 -14.747  39.233  1.00  0.00      ION CL
END
```

APPENDIX-I: EXCERPT OF LINEAR PSF (PROTEIN STRUCTURE FILE)

(This file contains nearly 3000 lines and provides complete bond information for the whole molecular system)

```
PSF

      11 !NTITLE
 REMARKS original generated structure x-plor psf file
 REMARKS topology C:/Program Files/University of
Illinois/VMD/plugins/noarch/tcl/autoionize1.2/ions.top
 REMARKS topology C:/Program
 REMARKS segment N1 { first ; last ; auto none  }
 REMARKS segment N2 { first ; last ; auto none  }
 REMARKS segment N3 { first ; last ; auto none  }
 REMARKS segment N4 { first ; last ; auto none  }
 REMARKS segment WT1 { first NONE; last NONE; auto none  }
 REMARKS segment WT2 { first NONE; last NONE; auto none  }
 REMARKS segment WT3 { first NONE; last NONE; auto none  }
 REMARKS segment ION { first NONE; last NONE; auto none  }

   16386 !NATOM
       1 N1   1    THY  C4'  CN7    0.160000       12.0107           0
       2 N1   1    THY  H4'  HN7    0.090000        1.0079           0
       3 N1   1    THY  O4'  ON6   -0.500000       15.9994           0
       4 N1   1    THY  C1'  CN7B   0.160000       12.0107           0
       5 N1   1    THY  H1'  HN7    0.090000        1.0079           0
       6 N1   1    THY  C2'  CN8   -0.180000       12.0107           0
       7 N1   1    THY  H2'  HN8    0.090000        1.0079           0
       8 N1   1    THY  H2'' HN8    0.090000        1.0079           0
       9 N1   1    THY  H5T  HN5    0.430000        1.0079           0
      10 N1   1    THY  O5'  ON5   -0.660000       15.9994           0
      11 N1   1    THY  C5'  CN8B   0.050000       12.0107           0
      12 N1   1    THY  H5'  HN8    0.090000        1.0079           0
      13 N1   1    THY  H5'' HN8    0.090000        1.0079           0
      14 N1   1    THY  N1   NN2B  -0.340000       14.0067           0
      15 N1   1    THY  C6   CN3    0.170000       12.0107           0
      16 N1   1    THY  H6   HN3    0.170000        1.0079           0
      17 N1   1    THY  C2   CN1T   0.510000       12.0107           0
      18 N1   1    THY  O2   ON1   -0.410000       15.9994           0
      19 N1   1    THY  N3   NN2U  -0.460000       14.0067           0
      20 N1   1    THY  H3   HN2    0.360000        1.0079           0
                                       .
                                       .
                                       .
    2015    2014    2016    2011    2018    2016    2020    2019
    2045    2042    2047    2046    2049    2047    2051    2050
    2052    2043    2049    2051    2077    2074    2079    2078
    2081    2079    2083    2082    2084    2075    2081    2083
    2109    2106    2111    2110    2113    2111    2115    2114
    2116    2107    2113    2115    2149    2139    2143    2148
    2149    2148    2150    2151    2173    2170    2175    2174
```

```
2177    2175    2179    2178    2180    2171    2177    2179
2213    2203    2207    2212    2213    2212    2214    2215
2236    2237    2240    2238    2240    2239    2241    2236
2243    2241    2245    2244    2269    2270    2273    2271
2273    2272    2274    2269    2276    2274    2278    2277
2305    2300    2307    2306    2309    2303    2307    2308
2309    2308    2310    2311    2332    2333    2336    2334
2336    2335    2337    2332    2339    2337    2341    2340
2366    2363    2368    2367    2370    2368    2372    2371
2373    2364    2370    2372    2398    2395    2400    2399
2402    2400    2404    2403    2405    2396    2402    2404
2430    2427    2432    2431    2434    2432    2436    2435
2437    2428    2434    2436    2470    2460    2464    2469
2470    2469    2471    2472    2494    2491    2496    2495
2498    2496    2500    2499    2501    2492    2498    2500
2538    2528    2532    2537    2538    2537    2539    2540

   0 !NDON: donors


   0 !NACC: acceptors


   0 !NNB

   0       0       0       0       0       0       0       0
   0       0       0       0       0       0       0       0
                                 END
```

LIST OF REFERENCES

Adler, B. J., & Wainwright, T. E. (1957). Phase transition for a hard sphere system. *Journal of Chemical Physics, 27*, 1208-1209.

Baldwin, G. S., Brooks, N. J., Robson, R. E., Wynveen, A., Goldar, A., Leikin, S., et al. (2008). DNA Double Helices Recognize Mutual Sequence Homology in a Protein Free Environment. *J. Phys. Chem. B, 112*(4), 1060-1064.

Becker, O. M. (2001). *Computational biochemistry and biophysics*: CRC Press, 2001.

Beveridge, D. L., Barreiro, G., Byun, K. S., Case, D. A., III, T. E. C., Dixit, S. B., et al. (2004). Molecular Dynamics Simulations of the 136 Unique Tetranucleotide Sequences of DNA Oligonucleotides. I. Research Design and Results on d(CpG) Steps. *Biophysical Journal, 87*, 3799-3813.

Bhandarkar, M., Brunner, R., Chipot, C., Dalke, A., Dixit, S., Grayson, P., et al. (2008). NAMD User's Guide
Version 2.6: Theoretical Biophysics Group University of Illionis and Beckman Institute.

Bragg, W. L. (1914). The Structure of Some Crystals as Indicated by Their Diffraction of X-rays. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character, 89*(610), 248-277.

Carlton, P. M., Cowan, C. R., & Cande, W. Z. (2003). Directed Motion of Telomeres in the Formation of the Meiotic Bouquet Revealed by Time Course and Simulation Analysis. *Molecular Biology of the Cell, 14*(7), 2832-2843.

Cramer, C. J. (2004). *Essentials of computational chemistry : theories and models*. Chichester, West Sussex, England; Hoboken, NJ: Wiley.

Crick, F. H. (1965). Recent Research in Molecular Biology: Introduction. *British Medical Bulletin, 21*, 183-186.

Darden, T., Perera, L., Li, L., & Pdersen, L. (1999). New tricks for modelers from the crystallography toolkit: the particle mesh Ewald algorithm and its use in nucleic acid simulations. *Structure*(7), R55-R60.

Derhaag, D. (1996). Recombinant DNA Labs: Basic Tools For The Molecular Biologist, *General Atomics Sciences Education Foundation Web Site*: General Atomics.

GraphPadSoftware. (2011). Quick Calcs Analyze a 2 X 2 Contingency Table. In H. Motulsky (Ed.), *Quick Calcs* (Vol. 2011): Graph Pad Software.

Gunawardena, S., & Rykowski, M. C. (2000). Direct evidence for interphase chromosome movement during the mid-blastula transition in Drosophila *Current Biology, 10*(5), 285-288.

Guvench, O., & Alexander D. MacKerell, J. (2008). Comparison of Protein Force Fields for Molecular Dynamics Simulations. *Methods in Molecular Biology, 443*, 63-88.

Hartley, H. O. (1949). TESTS OF SIGNIFICANCE IN HARMONIC ANALYSIS. *Biometrika, 36*(1-2), 194-201.

Hendrickson, J. B. (1961). Molecular Geometry. I. Machine Computation of the Common Rings. *Journal of American Chemistry Society, 83*(22), 4537-4547.

Isgro, T., Phillips, J., Sotomayor, M., & Villa, E. (2007). NAMD Tutorial - Windows Version: University of Illionis at Urbana-Champaign NIH Resource4 for Macromolecular Modelling and Bioinformatics.

Law, A. M., & Kelton, W. D. (2000). *Simulation Modeling and Analysis* (Third Edition ed.).

Leach, A. R. (2001). *Molecular Modelling - Principles and Applications*.

Ludford, P. J. *One Way Repeated Measures ANOVA*. Retrieved May 20, 2011, 2011, from http://www-users.cs.umn.edu/~ludford/Stat_Guide/repeat_meas_ANOVA.htm

Marshall, W. F., Straight, S., Marko, J. F., Swedlow, J., Dermburg, A., Belmont, A., et al. (1997). Interphase chromosomes undergo constrained diffusional motion in living cells. *Current Biology, 7*, 930-939 939.

Mathworks. (2009). *Glossary*. Retrieved July 11, 2009, from http://www.mathworks.com/access/helpdesk/help/toolbox/econ/index.html?/access/helpdesk/help/toolbox/econ/f6-1001427.html&http://www.google.com/search?hl=en&q=define%3Abrownian+motion&aq=f&oq=&aqi=

McCrum-Gardner, E. (2008). Which is the correct statistical test to use? *British Journal of Oral and Maxillofacial Surgery, 46*, 38-41.

McKee, B. D. (2004). Homologous pairing and chromosome dynamics in meiosis and mitosis. *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression, 1677*(1-3), 165-180.

Mercer, D. C. (2001). Data Decimation. What Do I Do?, *Prosig Noise & Vibration Measurement Blog*: http://blog.prosig.com/2001/06/06/data-decimation-what-do-i-do/.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. *Journal of Chemical Physics, 21*, 1087-1092.

Moler, C. (2004). MATLAB® 7.0 (Release 14): MathWorks.

Morozov, A. V., Kortemme, T., Tsemekhman, K., & Baker, D. (2004). Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum mechanical calculations. *Proceedings of the National Academy of Sciences of the United States of America, 101*(18), 6946-6951.

Muller, H. J. (1927). Artificial Transmutation of the Gene. *Science, 66*(1699), 84-87.

Muller, H. J. (1937). Physics in the Attack on the Fundamental Problems of Genetics. *The Scientific Monthly, 44*(3), 210-214.

Pauling, L., Corey, R. B., & Branson, H. R. (1951). The Structure of Proteins: Two Hydrogen-Bonded Helical Configurations of the Polypeptide Chain. *Proceedings of the National Academy of Sciences of the United States of America,, 37*, 205-211.

Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., et al. (2005). Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry, 26*(16), 1781-1802.

Pinchuk, A. O., & Vysotskii, V. I. (2001). Long-range intermolecular interaction between broken DNA fragments. *Physical Review E, 63*(3), 031904.

Pitman, J. W. (2003). Basic Properties of Brownian Motion, *Probability Theory Spring 2003 Lecture 15*. Berkeley.

Preacher, K. J. (2011). An interactive calculation tool for chi-square tests of goodness of fit and independence: University of Kansas.

Schlecht, M. F. (1997). *Molecular Modeling on the PC*.

Schrödinger, E. (1944). *WHAT IS LIFE? The Physical Aspect of the Living Cell Based on the lectures delivered under the auspices of the Dublin Institute for Advanced Studies at Trinity College, Dublin, in February 1943*.

Science, U. S. D. O. E. O. o. (2008). *Potential Benefits of the Human Genome Project Research*. Retrieved January, 2009, from http://www.ornl.gov/sci/techresources/Human_Genome/project/benefits.shtml

Stone, A. J. (2000). *The Theory of Intermolecular Forces*: Oxford University Press.

Stroud, J. (2006). *Automated Nucleic Acid Builder*. Retrieved December 20, 2008, from http://structure.usc.edu/make-na/server.html

Sybenga, J. (1999). What makes homologous chromosomes find each other in meiosis? A review and an hypothesis. *Chromosoma Focus, 108*, 209-219.

Szabo, A., & Ostlund, N. S. (1996). *Modern quantum chemistry : introduction to advanced electronic structure theory*. Mineola, N.Y.: Dover Publications.

Thibos, L. N. (2003). Fourier Analysis for Beginners (3rd ed.): Visual Sciences Group.

UCLA: Academic Technology Services, S. C. G. (2011). *Introduction to SAS*. Retrieved May 18, 2011, 2011, from http://www.ats.ucla.edu/stat/sas/whatstat/whatstat.htm

Wiberg, K. B. (1965). A Scheme for Strain Energy Minimization. Application to the Cycloalkanes. *Journal of American Chemistry Society, 87*(5), 1070-1078.

Widlund, H. R., Kuduvalli, P. N., Bengtsson, M., Cao, H., Tullius, T. D., & Kubista, M. (1999). Nucleosome Structural Features and Intrinsic Properties of the TATAAACGCC Repeat Sequence. *Journal of Biological Chemistry, 274*(45), 31847-31852.

Williams, D. (2004). Understanding FFT Windows: LDS Inc.

Zickler, D. (2006). From early homologue recognition to synaptonemal complex formation. *Chromosoma Focus, 115*, 158-174.

Zickler, D., & Kleckner, N. (1998). The leptotene-zygotene transition of meiosis. *Annual Review of Genetics, 32*, 619(617).