OUTLINE:

Paragraph 1-2: Modeling and its importance in understanding our past, present, and future situation/role in the universe.

Paragraph 3: Introduce the main idea of project. Introduce Brownian motion and discuss why it is a good approach to modeling complex systems.

Paragraph 4: Introduce Wasserstein Distance, discuss its use to real world data, how we are going to use.

Paragraph 5: How I became interested in these problems.

Paragraph 6: What tools I will need.

Paragraph 7: Closing Statements, ramifications of this research.

Modeling complex systems, from the growth of cancer cells and bacteria to the evolution of a species, is critical to our understanding of the physical world and our place in it. To comprehend humanity’s origin, it is essential to accurately model the dynamics driving the creation of life, and to comprehend our future, it is essential we have accurate predictive models. Such modeling is complicated when the dynamics behind such complex systems is random, or stochastic.

Most current models describing biological processes are deterministic, meaning they do not reflect any randomness in the system. This approach fails to grasp the truly random nature of the physical world. From the unpredictable way cancer divides to the unpredictable way genes split, randomness is a constant in the biological world. Deterministic approaches to modeling complex systems often produce inaccurate predictions, particularly over long periods of time. Dr. Motsch, a professor of applied biology and statistics in the School of Mathematical and Statistical Sciences, and I will study the same complex systems with the use of Brownian motion. Brownian motion, originally studied by the botanist Robert Brown in the 19th century and popularized by Einstein in the early 20th century, describes motion that behaves in a random way. If a particle suspended in water were acted upon randomly by other particles, its behavior would be described as Brownian. A variable is said to be Brownian if its derivative is a normally distributed random variable. My research with Dr. Motsch will seek to demonstrate that Brownian motion is a more accurate approach to modeling stochastic complex systems.

In order to quantify the accuracy of our approach, we will be using a tool in mathematics called the Wasserstein Distance, which quantifies the distance between functions or data sets. We will collaborate with a biologist from ISTB to obtain data on cancer growth in a mouse. With this data, we will use the Wasserstein distance to compute the distance between the true biological data, the predictions made by the Brownian motion approach, and the predictions made by the differential equations. We seek to understand how the number of particles used in the Brownian motion approach affects the accuracy of our approximation of the biological data. More specifically, we will establish convergence rates for the Brownian motion solution to converge to the real data as a function of the number of particles used.

Dr. Motsch and I are well positioned to complete this project. Dr. Motsch received his PhD in applied mathematics with an emphasis in animal cognition and mathematical biology, and works at the intersection of computational math, biology, and statistics. He has overseen undergraduate and graduate theses in flocking behavior, applied statistics, and offers the biological perspective necessary for this project. He has worked previously in the Wasserstein Distance and will provide the background in the WD to apply it to our research.

During a summer research project[[1]](#footnote-1), completed under Dr. Anne Gelb in 2014, I studied the use of statistical algorithms to improve Fourier edge detection in the presence of corrupted data. Specifically, Dr. Gelb and I derived an algorithm for reconstructing MRI images that is 166% more accurate than previous methods available to the medical community. In this project I developed an interest in using statistical tools to study relevant problems in biology. During the summer of 2016 I participated in two summer research programs that provided me with the interest and background necessary to successfully complete this project. My first project was at the University of California, Santa Barbara where I took a series of courses on differential equations in random media, biological modeling of living systems, and statistical approximations of biological processes. Through conversations with professors and graduate students from Princeton, Stanford, UC Berkeley and others, I learned about cutting-edge research at the intersection of biology and statistics. In my second project, I was selected with three other undergraduates from over 300 to participate in a research experience at San Diego State University. There I studied the convergence behavior of statistical algorithms used in linear regression models. Specifically, my advisor and I proposed the first algorithm for concrete convergence bounds for the error of a Gibbs sampler in a Bayesian linear regression model. This project will be submitted for publication in October[[2]](#footnote-2). This project taught me the tools to complete statistical convergence analysis, and I will use similar tools to study the convergence behavior of the approximation to complex systems using Brownian motion.

This project requires an advanced grasp on real analysis, probability theory, and numerical analysis. In order to complete this project successfully, I have enrolled in graduate courses in real analysis, distribution theory, and stochastic processes. In these courses I will obtain the tools necessary to read papers relating to my project as well as hold informed discussions with Dr. Motsch on our research. I have taken a course on numerical methods with Dr. Motsch, and I will use my textbook as a reference during my research. In the Spring I will finish the graduate sequence in real analysis and probability theory.

Developing accurate statistical algorithms for understanding stochastic complex systems is one of the largest problems in modern biological sciences, and the opportunity to study this type of problem has been my main focus during my undergraduate career. In addition to contributing to the field of computational biology, I hope to galvanize theoretical mathematics and statistics to solve problems of interest in applied fields, leading to more interaction and discussion between the previously distant groups of researchers. I believe, for instance, that the Wasserstein distance can be successfully applied to problems in computer science and chemistry. In addition to significantly adding to my knowledge in statistics, computational mathematics, I hope this research project will start a larger discourse on the use of statistical algorithms, specifically Brownian motion, in biology, chemistry, and other applied fields. To do this, I will do my undergraduate thesis on this project with Dr. Motsch.

The Origins Project is a first step towards galvanizing an interest among mathematicians and statisticians in using theoretical tools to study problems of interest in biology. With the Origins Project’s Undergraduate Research Scholarship, I will have the opportunity to demonstrate the utility of mathematics and statistics in computational biology. By improving the accuracy of approximations in biology, which studies the development of infectious diseases and our place in the universe, my research will shed light on the origin and future of humanity.

1. *A Statistical Framework for Detecting Edges from Noisy Fourier Data Using Multiple Concentration Factors,* Shane Lubold, Anne Gelb [↑](#footnote-ref-1)
2. *Non-asymptotic Bounds for the Root Mean Square Error in MCMC Algorithms,* Shane Lubold, Ethan Gwaltney, Jorge Carlos Román. [↑](#footnote-ref-2)