Research Statement

In order for students to grow and develop in a research lab, they require freedom to make mistakes. One of the main purposes of mentoring undergraduates is to help them develop into excellent scientists, even if they do not publish their work. Therefore, I expect students to become personally invested in their projects and brainstorm solutions to difficulties that may arise. Although students initially struggle with the uncertainty of research, they quickly realize that their publications and success in the lab are determined by their personal diligence to the project. This mentoring technique is difficult to implement because it requires dozens of projects to simultaneously be progressing in the lab in order to give new undergraduate researchers the space and time necessary to realize that they must fully invest in their projects in order to be successful. I have a track record of helping undergraduates to reach their full potential. I have directly mentored 31 undergraduates who have presented 15 posters at scientific conferences. Since some of those undergraduates were on multiple projects, 32 undergraduate poster coauthorships have resulted from work that I have directly mentored. Additionally, seven publications include undergraduate work, with 12 undergraduate coauthorships or undergraduate first authorships. I anticipate that we will soon publish 10 additional manuscripts with 20 undergraduate authorships from preprint manuscripts and work that has already been completed in the lab.

As a bioinformaticist, I specialize in algorithmic design and development with an emphasis in machine learning. I have a passion for research, and I have applied my skills to complete 20 publications, preprints, or draft manuscripts that are currently being reviewed by external collaborators (13 of which I am first author). Currently, I directly mentor 13 undergraduate bioinformaticists and one ¾ time bioinformaticist, although we are in the processes of hiring two additional ¾ time bioinformaticists and several additional undergraduates. I established an online project management tool using Microsoft Teams that enables me to mentor students even when campus is closed. Therefore, I anticipate that our projects will continue to progress during the COVID-19 pandemic with few interruptions.

My dissertation primarily explored the evolutionary conservation of codon usage and coding sequence length biases across the Tree of Life. Codon usage dynamics directly affect the translational efficiency of a gene because a limited supply of cognate tRNA anticodons are present in a cell, and it is less efficient for non-cognate tRNA to translate a gene. Additionally, certain codon conformations are more readily translated or transcribed. We explored biases in codon aversion and codon pairing across all domains of life in order to identify a potential phylogenetic signal in a novel character class. Codon aversion occurs when a codon is not used in a gene. Codon pairing occurs when codons that encode the same amino acid are located within a ribosomal window and the tRNA is recharged with an amino acid before it diffuses from the ribosome. In 2017, we discovered that codon aversion is phylogenetically conserved in tetrapods. Later, we showed that it was conserved in all domains of life. Likewise, we found that codon pairing is also phylogenetically conserved, and we developed a method to evaluate controversial node placement on the Tree of Life using codon biases. We then applied these techniques to human populations, and we showed that these biases alone are sufficient to assess population stratification. We developed a web server (https://cubap.byu.edu) that allows researchers to query various codon usage biases present in humans using the 1000 Genomes Project. We are currently assessing the extent to which synonymous variants affect codon usage dynamics in disease-specific databases from the Alzheimer's Disease Neuroimaging Initiative and the Alzheimer's Disease Genetics Consortium.

Another often-overlooked aspect of codon usage bias occurs when ramps of rare slowly-translated codons concentrate at the beginnings of gene sequences. Ramp sequences also coincide with fewer hydrogen bonds (i.e., lower GC content), which decreases the cost of unwinding dsDNA during transcription and increases overall transcriptional efficiency. These ramp sequences are predicted to evenly space polymerases or ribosomes and prevent downstream collisions, which increases the overall efficiency of transcription or translation. We published ExtRamp in *Nucleic Acids Research* in 2019, which is the first algorithm to systematically identify ramp sequences across a genome. Since then, we have explored the effects of ramp sequences on disease genetics.

We analyzed 13 exonic variants spanning 12 genes that are either implicated in Alzheimer's disease or are in linkage disequilibrium with a variant that is implicated in Alzheimer's disease. We found that six of the 12 genes contained a predicted ramp sequence, compared to 14% of human genes that are predicted to contain a ramp sequence. We show that synonymous variant *rs2405442:T:C* in *PILRA* (Paired Immunoglobin Like Type 2 Receptor Alpha) destroys a predicted ramp sequence. Our biological validations of this variant using qPCR show that this synonymous variant alone decreases the transcriptional efficiency of *PILRA* (p-value=$4.45x10^{-9}$), indicating that synonymous codon choice might directly affect disease by changing mRNA expression. We are currently exploring the effects of this variant on translational efficiency using an ELISA protein assay.

Additionally, I am actively investigating protein coevolution and its potential impact on Alzheimer's disease. We currently have a semi-functional web server (https://dca.byu.edu) that allows researchers to view areas of high coevolution between proteins. Protein coevolution occurs when residue pairs co-adapt due to conserved physical interactions or functional relationships and is detectible by calculating mutual information. Mutual information takes a multiple sequence alignment of an ortholog spanning at least 100 diverse species and calculates the evolutionary conservation of each residue. Previously, we used mutual information to prospectively identify interactions between two proteins in *Arabidopsis thaliana* that were later biologically validated, as well as to validate protein interactions in *Escherichia coli* that were first identified in the wet lab. Until recently, calculating pairwise coevolution for all proteins was computationally intractable. However, we developed a novel mutual information algorithm that decreases resource utilization by over 90% and allows us to explore proteome-wide coevolution. Since ortholog identification is instrumental to calculating protein coevolution, we also explored novel techniques to detect orthologous genes. In 2018, we developed an ortholog identification algorithm, JustOrthologs, that decreases computational runtime by more than 96% compared to other common algorithms. We are actively developing JustOrthologs version 2, which will allow us to compute orthologs across more diverse taxonomic groups.

Recently, my research has primarily focused on Alzheimer's disease etiology and disease progression. I helped to annotate 809 whole mitochondrial genomes for the Alzheimer's Disease Neuroimaging Initiative (ADNI), and we used machine learning on the ADNI dataset to predict cognitive decline with ~90% accuracy using only blood microarray data. We also used pedigree analyses of the Utah Population Database to identify rare genetic variants that likely cause excess mortality from Alzheimer's disease. Similarly, pedigree analyses also identified lineage-specific rare variants associated with exceptional longevity. Additionally, we have assessed next-generation sequencing biases and synergetic relationships between variants associated with Alzheimer's disease. We are currently determining the extent to which Alzheimer's disease subtypes significantly affect disease trajectory. We are also actively developing a web server (https://prs.byu.edu) that allows researchers to calculate polygenic risk scores from a variety of genome-wide association studies. We anticipate that this resource will facilitate the adaptation of polygenic risk scores and Mendelian randomization in large human case-control cohorts.

Although these projects span a variety of topics, they collectively indicate a pattern of successfully challenging current paradigms. My past research shows that I am well-prepared to take a multi-pronged approach to analyzing Alzheimer's disease and evolutionary biases, as well as manage the undergraduate teams that actively develop these projects. I currently manage a diverse portfolio of projects tailored specifically toward student interests, and I welcome student input in research. I am a strong proponent of open science, and I enjoy assisting undergraduates learn how to conduct replicable and high-quality research. I have found that I am most successful as a researcher when I spend time teaching undergraduates how to be more independent and take personal responsibility for their projects. Although it takes more initial effort on my part to mentor undergraduates, after a few months I am able to have proficient students oversee projects of their own, which increases my overall productivity. I have a strong publication record and success managing undergraduate teams, and I anticipate continuing to mentor undergraduate research in the future.