Thesis Defense

Computer Science Master's Program

"Sequence and Structure: A Combined Framework for Classifying Convergent Protein Families"

By **Samuel Kaplan**

Abstract:

Climate change has accelerated the need for carbon capture research. Carbonic anhydrase (CA), an enzyme present across Archaea, Bacteria, and Eukarya, catalyzes the hydration of CO2. However, the CA superfamily poses a complex computational challenge: nearly 150,000 poorly annotated sequences spanning 8 convergently evolved families must be classified to assess their catalytic potential for industrial applications.

Traditional bioinformatics tools fail to classify these proteins. They were designed for divergent evolution based on molecular clock theory. This principle breaks down when sequence similarity no longer indicates functional relatedness. Sequence-based methods, like Clustal Omega, can detect divergence within families but cannot distinguish intra-family variation from inter-family convergence at scale. Structure-based tools, such as Foldseek and TM-Align, detect functional similarity but may confuse convergent active sites across evolutionarily distinct families.

This thesis presents an iterative multi-modal framework combining sequence and structural analysis. Using diversified HMM profiles from manually curated consensus sequences, we annotate sequences with HMMER. Low-confidence predictions are then validated using structure-based methods, resolving cases where sequence homology alone is insufficient. Our approach classified approximately 95% of CA sequences with high confidence, bridging the gap between evolutionary relatedness and functional convergence. This framework addresses the CA annotation bottleneck and provides a template for classifying other convergently evolved enzyme families where traditional homology-based methods fail.

Date: Thursday, December 11th, 2025
Time: 1:00 PM – 3:00 PM
Location: 14-238b
Committee: Dr. Anderson, Dr. Oza, Dr. Davidson
Zoom Link: https://calpoly.zoom.us/j/85441775794