

Presenting Researchers

Sophie Allen	4	Kevin Mueller	13
Pankaj Chouhan	5	Alexander Peterson	14
Maia Erbes	6	Sanjeeb Poudel	15
Jhamieka Greenwood	7	Daryn Sagel	16
Brandon Gusto	8	Hannah Squier	17
Kunal Kanawade	9	Liam White	18
Marzieh Khodaei	10	Kevin Zeigler	19
Young Huan Kim	11	Jingze Zhang	20
Kathryn Merritt	12		

Kevin Zeigler

Ph.D. in Computational Science

Advisor: Alan Lemmon

Draft Genome of the Upland Chorus Frog, *Pseudacris feriarum*

We performed the first de novo genome assembly of *P. feriarum*, the Upland Chorus Frog, a small Anuran in the family Hylidae that can primarily be found throughout the Eastern US. These populations are likely undergoing incipient speciation based on female call preference.

In our testing, we generated contigs from PacBio Sequel II continuous long reads using the open-source genome assembler Canu. We scaffolded contigs using long range chromosome linking information, Dovetail Chicago and HiC, with open-source software SALSA 2. The assembled genome of *P. feriarum* represents an important tool in pinpointing exactly which genes drive female call preference and therefore speciation.

Future work will entail using the newly assembled genome in a genome wide association study to narrow down which specific genes are responsible for speciation in specific populations of *P. feriarum*.



Figure Caption: Visualization of chromosome capture linking information. The color on the heat bar is the interaction frequency; the x and y axis represent the position/scaffold in the genome assembly (scaffolds are not sorted). On the diagonals ($x = y$) the interaction frequency is the highest. Within scaffolds (grey lines) the frequency is significantly higher than the surrounding areas.

Syntactical Graph Neural Network for Authorship Attribution

Authorship Attribution is the process of identifying the author of an anonymous document. It involves using statistical and computational methods to analyze extensive collections of text. Traditional studies focus on finding useful patterns in linguistic and stylistic features to tell them apart. They also rely extensively on feature engineering. Many studies also use neural networks, which led to some improvements. However, they only consider local syntactical features and fail to deal with long-range dependencies. Our research presents a novel architecture based on word connections within sentences via dependency trees and connections between sentences via the self-attention mechanism. The sentence's sequential order and syntactical structure enable our model to outperform the state-of-the-art. We also conduct a comprehensive ablation study to analyze the effect of different linguistic components, including word order, word length, word frequency, dependency type, and sentence order.

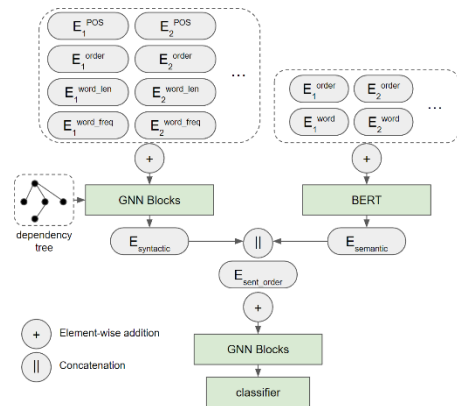


Figure. Architecture of the proposed syntactical graph neural network.

RESEARCH Abstracts & Graphics

Modeling Neural Circuits to Understand Incipient Speciation in Chorus Frogs

The current study probes at the genetic drivers of reproductive isolation in the chorus frog (*Pseudacris feriarum*) through analyzing the variation in female preference and male mating calls. This variation in male acoustic signaling is observed in *P. feriarum* allopatric populations, which have diverged due to geographic isolation, and in *P. feriarum* sympatric populations, which have diverged due to an interaction with another species (*P. nigrita*). This divergence is believed to be driven by changes in the gene expression of neurons in the anurans auditory midbrain that control mating behavior.

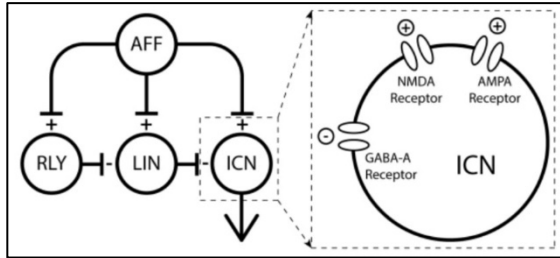


Figure 1. The neural circuit thought to allow chorus frogs (*Pseudacris*) to choose mates based on pulsed calls. The afferent neuron (AFF) sends signals to a relay neuron (RLY), a long-interval neuron (LIN), and an interval counting neuron (ICN). Through dis-inhibition, this circuit can be tuned to fire downstream only when the calls containing a range of pulse-rate and pulse number are heard by female chorus frogs

In order to understand the divergence of reproductive behaviors in *P. feriarum* populations, we modeled the neural circuitry of two allopatric (ancestral) and sympatric (diverged) populations. The computational model used was created by Naud et al., and integrated the three neurons that were modeled as leaky integrate-and-fire units, where each unit received excitatory and inhibitory inputs. The model was implemented and parameterized for each population based on previously obtained empirical preference data.

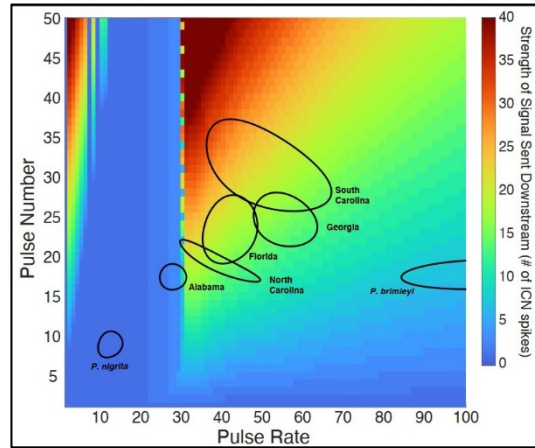


Figure 2. The response of the neural circuit to calls of different properties (pulse rate and pulse number) in the Florida population of *Pseudacris feriarum*. Ellipses show 50% confidence envelopes for the call distributions of different populations of *P. feriarum*. Note that energetic cost constraints prevent male calls from evolving to match areas of strongest female preference (i.e. red). Nonetheless, the tuning of the neural circuit allows females to avoid mating with males from other species (*P. nigrita* and *P. brimleyi*)

A Novel Approach for Adaptive Skeleton Toolpath Generation

Industry 4.0 is revolutionizing manufacturing through the integration of automation and real-time data sharing in cyber-physical systems. At the forefront of this revolution is large-format additive manufacturing. In large-format printing, parts are often designed to be an even number of beads wide to produce a completely dense part. However, voids can still arise. This is often due to the part not being an even number of bead widths wide in some areas, or in geometry containing acute angles, as the process of generating closed contours cannot completely fill the space. Voids can be tolerated in smaller models,

but in large-format additive manufacturing they may cause mechanical defects. To fill these voids, open loop paths called skeletons are often used, but they are typically limited by the physical constraints defined in the slicing software. To address this, researchers at Oak Ridge National Laboratory have extended skeleton toolpaths via an adaptive methodology. These adaptive skeletons were found to better fill void spaces through manipulation of physical parameters of the build process and were calculated as part of the slicing process.

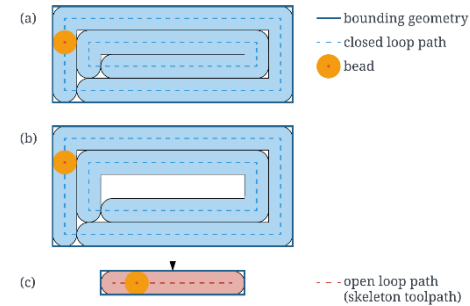
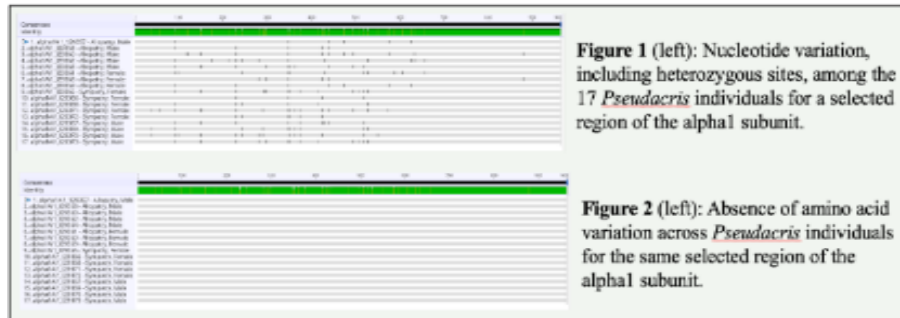


Figure above: (a) A rectangular geometry that is four beads wide is completely filled by two closed loop paths. (b) A rectangular geometry that is five beads wide is partially filled by two closed loop paths and contains a void that cannot be filled by an additional closed loop path of the same bead width. (c) The void featured in (b) is filled by an open loop path (skeleton toolpath).

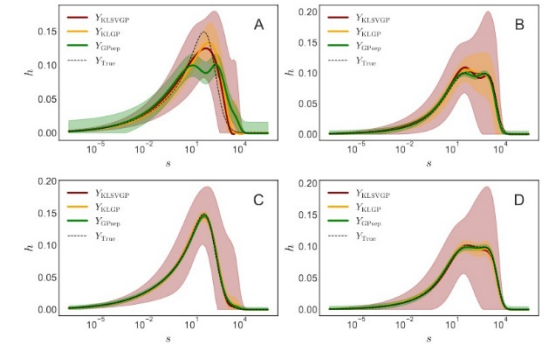
The Evolution of the GABA Receptor in Pseudacris and its Contribution to Speciation

This project analyzes candidate Pseudacris (chorus frog) genes associated with the neurological GABAA receptor in order to identify protein variation that could potentially cause differences in mate preferences across sex and geographical variables. A mathematical model suggests that female chorus frogs may discriminate between pulsed calls by utilizing neuron systems that communicate through neurotransmitters (Naudetal., 2015). A difference in GABA receptors or their protein subunits among populations could correspond to differences in female preferences to male calls, leading to speciation within Pseudacris populations. In this first study, we focus specifically on the alpha1 subunit of the GABAA receptor to determine and analyze amino acid variation in that region.



Efficiently Calculating the RPA Energy Difference between Two Similar Systems Based on Atom-Centered Correlated Sampling

In the polymer industry, molecular simulations are used to optimize product quality, increase revenue, and improve manufacturing processes. Molecular simulation models of polymer rheology take the molecular structure of a polymer mixture as input and yield the linear or nonlinear viscoelasticity of the mixture as output. Different molecular simulation models offer different tradeoffs between accuracy and speed. For some applications, the computational cost of molecular simulations can be prohibitive; this is where surrogate models can help. Surrogate models mimic the input-output behavior of the underlying molecular simulation to reasonable precision at a fraction of the computational cost.



This work explores a surrogate model based on a blend of Gaussian processes regression (GPR) with Karhunen-Loève expansion (KL). First, we developed an empirical covariance matrix from observations. Thereafter, an eigendecomposition of the empirical covariance matrix is performed to obtain principal modes and corresponding eigenvalue. A low-dimensional covariance matrix representation is created via selecting a few important principal modes. Each selected principal mode is then fitted using a GPR. To enable batch training, we explored stochastic training of GPR. We systematically studied the proposed model on training sets of varying sizes. A comparison of the current work with our previous work based on the separable kernel is also provided.

Figure 1: Here we are comparing the three developed surrogates, GPsep - a GP surrogate based on a separable kernel, KLGP - a surrogate based on GP and Karhunen-Loève expansion (KL), KLSVGP - an extended version of KLGP with batch training enabled.

Temporal Analysis of Chorus Frog Vocalizations

Acoustic signaling is vitally important in sexual selection, species recognition, and speciation across many taxa, including everything from reptiles to birds and mammals. Of these many taxa, anurans are an excellent choice for acoustic analysis, because acoustic signaling is known to be the strongest determining factor in their mate selection. Temporal variation across call sequences is known to be a factor in sexual selection, however fine scale temporal variation, such as within a single calling sequence, has been under examined.

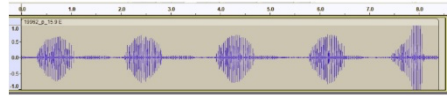


Figure 1: Oscillogram of five *P. feriarum* calls in Audacity.

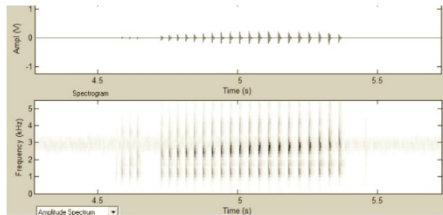


Figure 2: Oscillogram and spectrogram of a single *P. feriarum* call in Soundruler.

Pseudacris feriarum is an ideal system for this analysis because the spectral properties of their call sequences are known to show significant differences between allopatric and sympatric populations, resulting in dramatic effects on speciation.

In this study, I analyzed recordings of male *Pseudacris feriarum* vocalizations to determine how the acoustic characteristics of a call sequence changed from beginning to end.

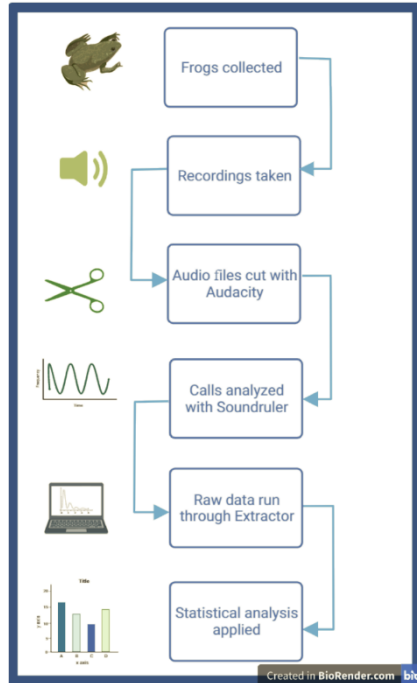
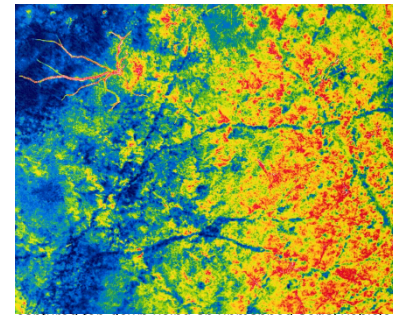


Figure 3. Process.

Data Collection for Wildland Fire: Isolating Fire and Plume Behavior

When fire propagates through a fuel bed, there are a variety of structures and behaviors we wish to better understand. Though fire occurs in diverse environments and may look significantly different depending on the region you are in, the physical laws that underlie fire behavior remain consistent. As such, we are able to study fire spread and plume behavior using data from a selection of fire scenarios. Two examples that are generalizable to multiple regions are (1) small-scale fire spread through a porous pine straw bed, which is common in nature, and (2) large-scale interactions between the fire, plume, forest canopy, and atmosphere. I have developed techniques that combine computer vision and



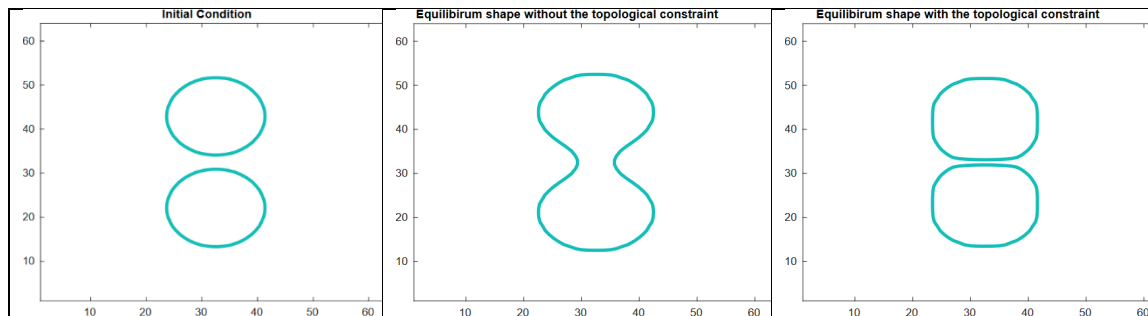
captured top-down using an unmanned aerial vehicle.

graph theory to track these behaviors. The data sets presented in this poster are from a combination of ground-based visual and infrared cameras, as well as aerial imagery capture by a small unmanned aerial vehicle. Due to the turbulent nature of fire, a statistical approach is used to analyze the data obtained through my method. This analysis reveals patterns behind the behavior of the fire and plume dynamics, heat transport, and fire-atmosphere coupling.

Topology-preserving Phase-field Modeling of Elastic Bending Energy

The phase-field method is an effective computational approach for solving interfacial problems and has been widely used in the modeling of conformation changes in biological lipid vesicles. The elastic bending energy of the vesicle formulated by phase-field dynamics is optimized using the standard gradient flow approach. Surface area and volume constraints imposed during the simulations give interesting shapes of the membranes. Most previous works, however, ignore the Gaussian curvature from the elastic bending energy, which allows the numerical simulations to automatically handle topological changes to the configurations of vesicle membranes. Our research recognizes that in some events (such as in the simulation of blood

cells), it might be important to preserve the topological information. In this work, we impose a topological constraint along with other constraints to calculate the equilibrium shape of membranes. We have been studying the constraint using the penalty formulation and using the discrete form of the formula to approximate the variation of topology with respect to the phase-field function. With our method, we have tested many cases, and it preserves the topology of the initial condition. Our future work will use the adaptive finite element method and explore various methods for imposing constraints, such as the method of Lagrange multipliers.



Adverse Effects on Erosion and Biogeochemical Fluxes Post-Fire

Soil is a simple substrate that we walk on every day. We do not think about its composition and importance in our environment. However, soil plays a vital role in various ecosystems and our environment. When we think about fire, it is a natural disturbance that humans have used as a land management tool for centuries. More specifically, a wildfire can generate a large amount of heat within a short period on the soil surface, causing significant biological, chemical, and hydrological effects (Bao et al., 2020). Furthermore, it leaves the soil exposed and easily erodible by precipitation or hardens the soil

surface causing poor drainage and erosion of loose sediments (USGS, 2017).

During the Spring 2022 Fire Dynamics Field School, we conducted experiments to study soil conditions and erosion rates after simulated precipitation for unburned and burned soil samples. We also considered the soil's land grade and texture, which both help to understand how the heat transfer process is affected. This work presents the analysis results conducted on water samples at the College of Engineering and the National High Magnetic Field Laboratory.

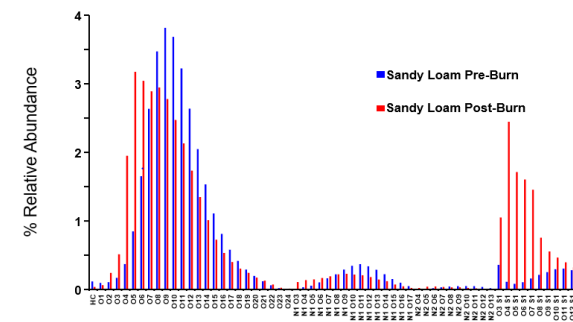


Figure 1. Heteroatom class distribution for acidic species derived from water-soluble species passed through sandy loam before and after burning.

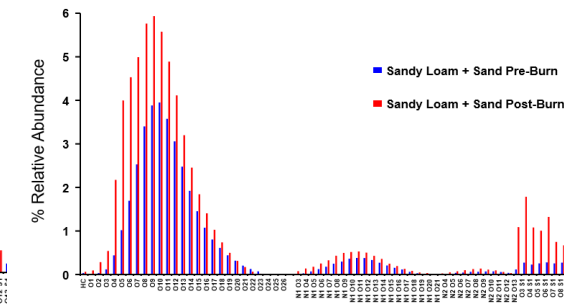


Figure 2. Heteroatom class distribution for acidic species derived from water-soluble species passed through sandy loam + sand before and after burning. The heteroatom class corresponds to species that only contain carbon and hydrogen.

Learning How Stars Explode: Supernovae, Turbulent Combustion, and Machine Learning

A novel, data-driven model of deflagration-to-detonation transition (DDT) is constructed for application to explosions of thermonuclear supernovae (SN Ia). The DDT mechanism has been suggested as the necessary physics process to obtain qualitative agreement between SN Ia observations and computational explosion models. This work builds upon a series of studies of turbulent combustion that develops during the final stages of the SN explosion. These studies suggest that DDT can occur in the turbulized flame of the white dwarf via the Zel'dovich reactivity gradient

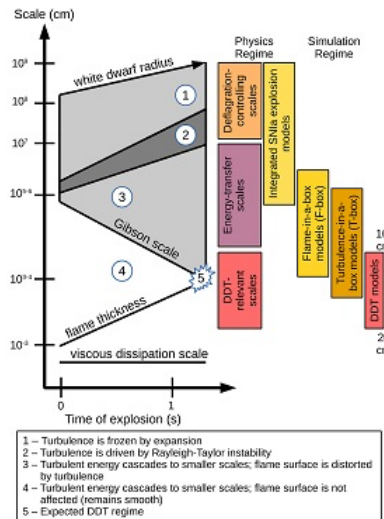


Figure 1: Spatial and temporal scales of the white dwarf during the explosion process. Physical regimes and their corresponding computational modeling regimes are highlighted. The spatial scales range from the deflagration-controlling scales on the order of the size of the white dwarf, down to the viscous dissipation scale. DDT occurs on scales slightly larger than the viscous dissipation scale.

mechanism when hotspots are formed. We construct a large database of direct numerical simulations that explore the parameter space of the Zel'dovich initiated detonation. We use this database to construct a neural network classifier for hotspots. The classifier is integrated into our supernova simulation code, FLASH/Proteus, and used as the basis for a subgrid-scale model for DDT. The classifier is evaluated both in the training environment and in reactive turbulence simulations to verify its accuracy in realistic conditions.

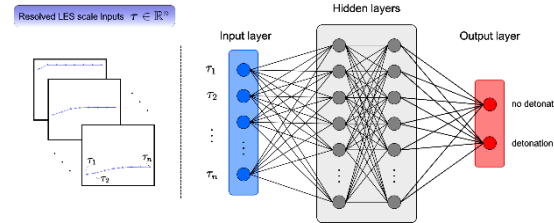


Figure 2: Overview of the neural network approach for one of the feature selection strategies. The network input is the induction time profile computed via DNS, but coarsened to the LES filter scale. The binary classification labels of 'no detonation' and 'detonation' are determined by the result of the corresponding DNS.

Utilizing Scientific Machine Learning in Oscillatory Shear Rheology Models

In rheology, modeling fluids by applying stress (especially non-Newtonian fluids) is important for glean information on fluid behavior. An efficient method of doing this is to read liquid- and solid-like (i.e. viscoelastic) properties on such fluids by applying oscillatory shear strain. In modeling the

resulting stress fields, some state-of-the-art rheological models are presented, and a Tensor-Basis Neural Network is utilized in order to model missing terms in the system of ODEs directly to mitigate limitations presented by conventional methods.

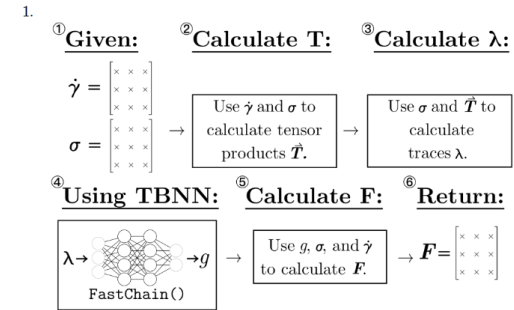


Figure 1: The basic structure of `tbnn()`. The tensor $\dot{\gamma}$ is equivalent to D in this function for calculating \hat{T} .

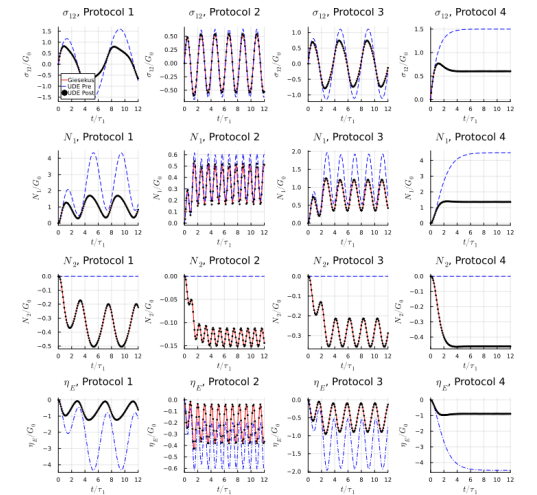


Figure 2: The resulting plot after training with 200 epochs per protocol (≈ 1600 epochs). The NN structure is $9 \rightarrow 32 \rightarrow 32 \rightarrow 9$ such that the activation function at each layer is `tanh`.

A Longitudinal MRI Preprocessing Pipeline of Grey Matter Atrophy for Modeling Alzheimer's Disease Progression

Alzheimer's disease is thought to be primarily characterized by emergence of amyloid beta ($A\beta$) plaques and neurofibrillary tangles from an abundance of tau, that spread among the brain inducing large amounts of neurodegeneration. One approach for modeling the spread of the disease is as a diffusion process of anatomical MRI-derived atrophy along a graph created from the brain's structural connectome or functional connectivity. However, this diffusion approach requires longitudinal time-series data that is often highly corrupted due to noise introduced in the standardization process when registering to a common space among subjects. In this work, we create a custom pipeline that includes: creating a longitudinal template with corresponding Jacobian determinant maps, segmenting the template into grey matter voxel probabilities, normalizing the template into a standardized space, partitioning the Jacobian determinant maps into regions, and creating an atrophy time series for all subjects with respect to a baseline scan. We demonstrate the effectiveness of our pipeline by applying it to a large cohort of subjects with varying baseline diagnosis from the ADNI database, and by comparing the resulting data with idealized synthetic data of the diffusion process.

Figure 1, Below: Regional Jacobian determinant time series of 116 regions derived from the AAL-atlas arranged by temporal lobe applied to two subjects from the AD and CN cohorts.

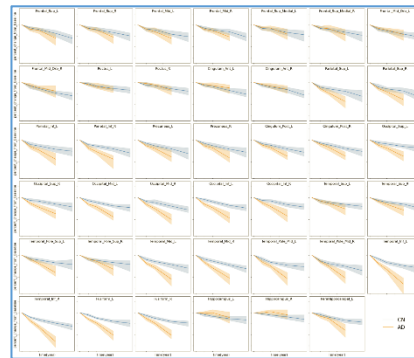
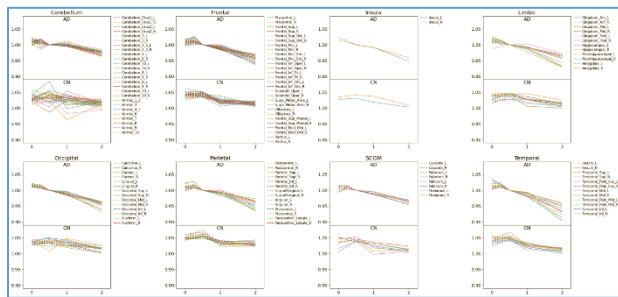


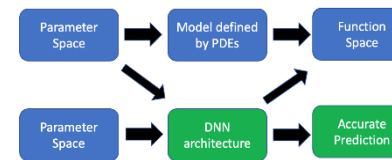
Figure 2, Above: Median time series for 41 important regions from the AAL-atlas of a percent atrophy from baseline measure for AD and CN subjects with corresponding 99% confidence intervals.

Approximation of Parameterized Partial Differential Equations Using Deep Learning Techniques

Physical phenomena are often described by complex partial differential equations. Optimizing for a particular parameter for a physical system can be computationally expensive due to repetitive computations. Deep learning techniques have worked well with data with large dimensions. Hence, we propose to use deep neural network architecture

to learn the mapping between the parameter and function spaces. In this research work, we would like to focus on formulating Deep neural learning high-dimensional functions with deep neural network architecture. This is ongoing research work, and more results are in progress.

Figure 1: Schematic for the approximation of parameterized partial differential equation.



Evolutionary Analysis Using Probabilistic Topic Modeling

Biological data classification and taxonomic identification have a significant role in evolutionary biology and bioinformatics. Most current approaches use a two-step procedure to classify biological data: (1) alignment of the biological sequences, (2) analysis of this alignment. Alignment-based sequence analysis might be troublesome. The alignment-free approaches are gaining more attention from the scientific community because of their ability to overcome the drawbacks of sequence alignment

techniques. In this project, we adopt the alignment-free technique of probabilistic topic modeling, a statistical approach based on document analysis techniques, to extract topic frequencies of the database, and then for evolutionary analysis based on topic frequencies, we base our development on CONTML (Gene Frequencies and Continuous Characters Maximum Likelihood method), a software module of the PHYLIP package by Felsenstein.

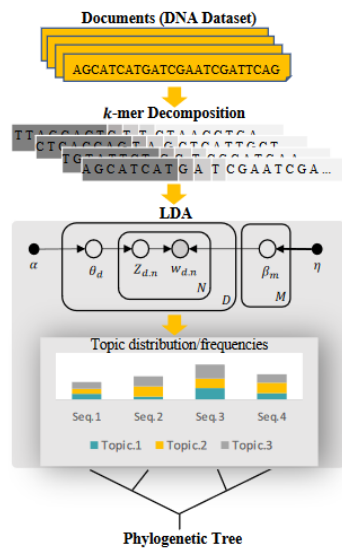


Figure 1. Workflow of topic modelling to generate topic frequency for documents of sequences, and the estimated phylogeny using obtained topic frequencies and CONTML.

Evaluating Genomic Variation

In a recently patented method, polymerase chain reaction is possible using only one primer. The method uses two common adapter sequences and a primer designed to match a repetitive element from a genome. Primers can be selected to target a particular number of loci. The focus of this study was to apply a computer program to several example genomes to identify possible candidate primers. The program used was written in java and took in a fasta file of the genome. The genomes used ranged from 0.23 Gb to 4.23 Gb and covered plants, amphibians, mammals, insects, and invertebrates. Candidate primers were selected based on several criteria such as length, how often the flanking regions could be mapped to the given genome, and the expected number of retrieved loci. Once candidate primers were returned by the computer program, they can be further analyzed for possible issues such as the possibility of binding with themselves or the adapter sequences. Since a maximum of 20 candidate

primers were selected from each genome, a selection of primers from each genome was chosen to be run through a second computer program. The second program returned specific qualities of the primer such as melting temperature, average distance between a corresponding locus and the closest mappable region, and the percent of retrieved loci that had a mappable region within 300 base pairs. Some of the primers have been tested using the recently patented method in lab, but further work is needed to demonstrate that the theoretical results match the experimental results across species. If the experimental results match the theoretical results across the wide variety of species used in this study, then this method could be applied for more cost-effective population genetics, forensics, and paternity tests. Due to the methods ability to target a specified number of loci, it also has the ability to be used in a wide variety of projects regardless of the number of loci needed.

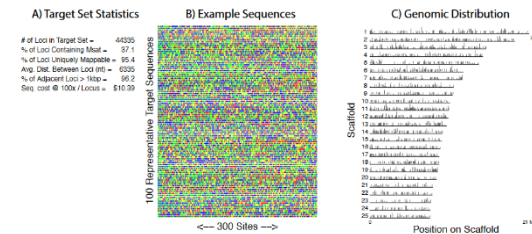


Figure 1, Left: Preliminary analyses are used to identify target sets with desirable properties. Here, a draft chorus frog genome (4.4 Gb) was analyzed to identify a target set containing ~45k loci. A) The following target set statistics are computed: 1) percent of loci containing one or more microsatellites (min repeats=8), 2) percent of loci uniquely mappable (locus considered mappable if it contains one or more 30-mers that occur only once in the genome), 3) average number of nucleotides between loci (a measure of marker density), 4) percent of adjacent loci > 1kb (a measure of locus independence), and 5) per-sample sequencing cost assuming paired-end 150bp Illumina sequencing on an S4 ow cell with coverage predicted to yield 100 read-pairs per locus. B) Target region properties are also assessed visually using 100 representative target sequences of length 300. This depiction does not show variation across samples, but instead shows base composition and presence of microsatellites for different loci from the same individual. C) Finally, the distribution of target loci is shown for the longest 25 scaffolds/chromosomes.