SAND2024-13351C

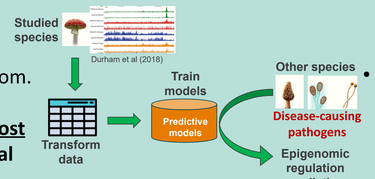
Sandia National Laboratories

MIXED CNN-ATTENTION MACHINE LEARNING MODEL FOR PREDICTING GENE REGULATORY RELATIONSHIPS ACROSS FUNGAL SPECIES AS A COMPUTATIONAL METHOD TOWARDS DEFENDING AGAINST EMERGING PATHOGENIC FUNGI

Laura Weinstock, Jenna Schambach, Anna Fisher, Cameron Kunstadt, Elizabeth Koning, Wittney Mays, and Raga Krishnakumar

Abstract

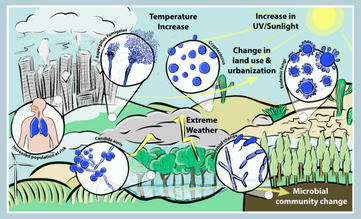
- The discovery of conserved epigenetic modification design rules that control gene expression would greatly improve the efficiency of therapeutic discovery and countermeasure development in preparation for threats spanning the fungal kingdom.
- **We need ways to understand evolving behavior across fungal species with speed, accuracy and cost effectiveness to take advantage of their beneficial properties and combat their harmful effects.**



- We uses ML to predict gene expression levels based on combinatorial epigenetic modification expression within and across fungal species.
- To support a) more accurate prediction of how fungal species change when modulating epigenetics and b) determine how to target critical genes and cellular functions as a means to identify targets for antifungal therapeutics.

Background

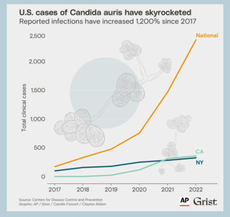
A changing world creates novel fungal properties and threats



Van Rhijn and Bromley. (2021).

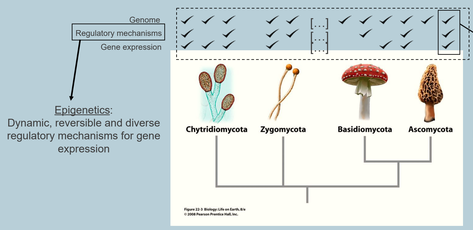
- Fungal pathogens present a challenge for public health, as they are diverse, largely uncharacterized, and challenging to treat
  - Current estimates for total fungal species are 2-11 million
  - Approximately 3000 fungi are in completed or ongoing sequencing projects
  - Only about 900 full fungal genomes have been released
- Drug resistance and climate-driven pressure exacerbate the issue
- There is a need for the rapid discovery, development, and efficient repurposing of countermeasures to combat fungal infection at scale

Fungal infections are on the rise



*Currently most prevalent in immunocompromised patients

Cross-species prediction of fungal biology is needed for understanding phenotypic changes at scale



Epigenetics:
Dynamic, reversible and diverse regulatory mechanisms for gene expression

Most current work:
Individual species, in-depth analysis

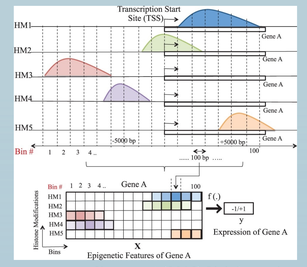
Missing:
Large, cross-species, cross-study analysis

Challenges:
Technical variability
Biological/functional diversity
Transforming data

Key questions:
1) **Are epigenomics predictive of gene activity in fungi?**
2) **Can information from one fungal species be predictive in another?**

- Epigenetic mechanisms are critical in antifungal drug resistance and are a rich source of drug targets
- Epigenetic mechanisms are crucial in regulating gene expression, and impact fungal growth and virulence
- Understanding epigenetics in fungi can help us develop better antifungals

Previous work: CNN for Epigenetic Prediction of Gene Expression



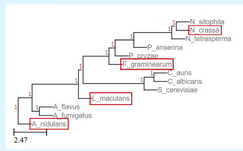
Singh *et al*. (2016) *Bioinformatics.*

- Recent efforts have sought to understand the relationship between gene sequence epigenetic modifications, and gene expression within fungal species through the use of machine learning and deep learning (ML/DL) methods
- The discovery of conserved epigenetic modification design rules and functional (gene expression) impacts would greatly improve the understanding of emerging pathogens and antifungal targets for diverse species

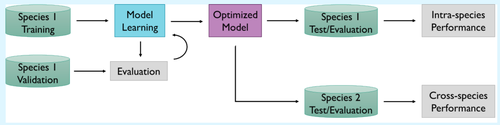
Results

Phylogenetic distance

- Species used in analysis span genetic distances
- Tree developed by custom phylogenetic algorithm

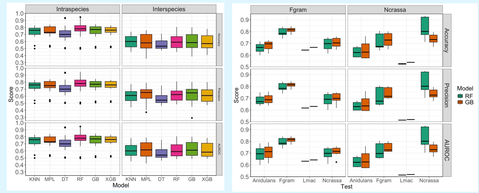


Machine learning training and evaluation strategy



- Epigenetic markers used as predictor for gene expression
- Intra and inter species prediction tasks

Shallow model results



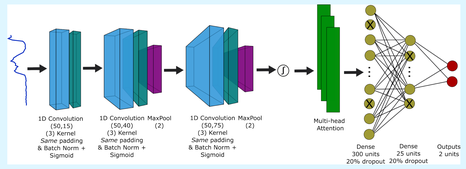
- Tree-based models outperform other shallow models on *N. crassa*
- Cross-species predictions using shallow learning models have below 60% accuracy

Legend:
Lmac = Leptosphaeria maculans;
Ncrassa = Neurospora crassa;
Fgram = Fusarium graminearum

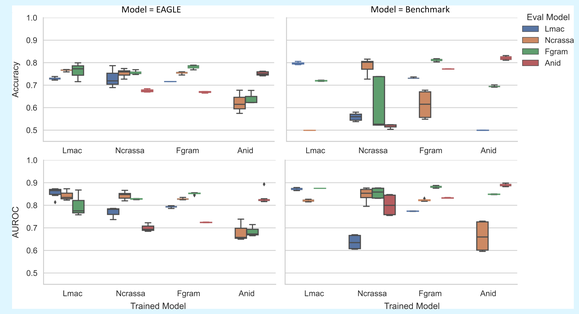
Deep learning model results

- Model output is 2-class prediction of gene expression
- Architecture stacks three 1D CNN layers, a multi-head attention layer, and two fully-connected layers

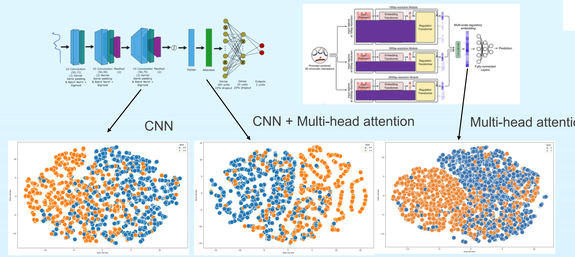
EAGLE Architecture



- Custom model outperforms benchmark model (modified Chromoformer) on cross-species predictions (75-80% accuracy)



- tSNE visualization of latent embeddings show distinct representations



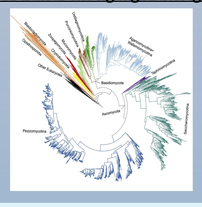
- Fraction of overlapping motifs discovered by HOMER for genomic regions identified by SHAP as important for EAGLE's prediction-making

	K4me3	K4me2	K27me3	K36me3
Top 50 motifs	0.38	0.54	0.46	0.36
Top 100 motifs	0.53	0.58	0.37	0.45
Right				
Top 50 motifs	0.38	0.58	0.46	0.34
Top 100 motifs	0.45	0.62	0.5	0.41
Wrong				
Top 50 motifs	0.46	0.52	0.5	0.36
Top 100 motifs	0.47	0.53	0.47	0.43

Conclusions

- Shallow models have limited predictive power for inter-species tasks
- Custom CNN + Attention model predicts intra- and cross-species gene expression based on epigenetic modifications signal profiles
- We confirm conservation of some epigenetic mechanisms across species
- Limitations in predictive capacity may be due to underlying biological constraints or limited data availability.
- Modular gene expression through understanding and predictive modeling of regulatory epigenetic modification rules is possible through machine learning and can support countermeasures for fungal pathogens.

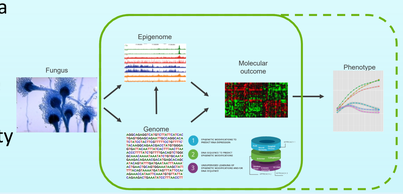
Potential for Prediction Across More and Emerging Pathogens



Future Work

- Building more robust predictive models by adding additional and multi-modal data from published and in-house sequencing work
- Continue exploration into adaptability of tools across more species
- Identify and experimentally validate epigenetic modification profiles that regulate gene-of-interest expression
- Predict what epigenomic mechanisms are likely to drive resistance or pathogenicity in novel fungal threats and how to target the epigenome to counter this

Prospective Goal: Predicting Phenotypes Of Fungal Species From Genomic And Epigenomic Information



U.S. DEPARTMENT OF ENERGY
National Nuclear Security Administration

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

SAND2024-13351C

Sandia National Laboratories