

United States Department of Agriculture
Project Initiation

Project Data for Accession Number 1035015

Project Title

DSFAS: Discovery of fungal CAZymes by leveraging protein 3D structures and AI

Project Details

Sponsoring Institution National Institute of Food and Agriculture	Project Status ACTIVE
Funding Source NIFA Non Formula	Grant Year 2027
Grant No. 2027-67021-47661	Cumulative Award Amt. \$585,000.00
Proposal No. 2025-11509	Multistate No. (N/A)
Project Start Date Nov 01, 2026	Project End Date Oct 31, 2029
Program Code [A1541] Food and Agriculture Cyberinformatics and Tools	

Project Director

Yin, Y.

Recipient Organization

BOARD OF REGENTS OF THE UNIVERSITY OF NEBRASKA
2200 VINE ST # 830861
LINCOLN, NEBRASKA 685032427

Performing Department

(N/A)

Non Technical Summary

The PD Yin is a Computational Biologist in the Department of Food Science and Technology (FDST) and Nebraska Food for Health Center (NFHC) of UNL. His long-term research goal is to develop new computational tools for Carbohydrate-Active Enzyme (CAZyme) annotation in microbial genomes. In the last decade, his team has developed impactful software tools and data infrastructures, collectively called the dbCAN tool suite, which has processed >500,000 user-submitted jobs to the dbCAN web server in the last 10 years. The objective in this project is to leverage protein 3D structures and artificial intelligence (AI) to further develop new tools for the dbCAN tool suite to assist the discovery of new CAZymes and associated gene clusters in fungal genomes. Major research gaps exist in that the current computational tools (including dbCAN) for CAZyme discovery: (1) are primarily designed for bacterial genomes, and (2) are heavily based on sequence similarity search, which limit their ability to discover novel fungal CAZymes. The innovation of this project is that new computational tools will be delivered to detect novel CAZyme

United States Department of Agriculture Project Initiation

families without sequence similarity but with structural similarity and semantic similarity to known CAZymes. The significance of this project is that it will undoubtedly facilitate the discovery of new fungi CAZymes that can be harnessed for the sustainable production of bioenergy, for improved feed efficiency in livestock, and for enhancing various aspects of the food industry. This project focusing on new fungal CAZyme discovery will have major impact in multiple USDA AFRI priority areas, such as Plant health and production and plant products; Bioenergy, natural resources, and environment.

Animal Health Component

Animal Health Component: 0%

Research Effort Categories

Basic	100%
Applied	0%
Developmental	0%

Classification

Knowledge Area (KA)	Subject of Investigation (SOI)	Field of Science (FOS)	Percent
212	4020	1080	100%

Knowledge Area

[212] Diseases and Nematodes Affecting Plants

Subject Of Investigation

[4020] Fungi (includes yeast)

Field Of Science

[1080] Genetics

Keywords

bioenergy, fungi, plant disease, protein structure, CAZyme, genomic context, protein language model

Goals / Objectives

The objective in this project is to leverage protein 3D structures and artificial intelligence (AI) to further develop new tools for the dbCAN tool suite to assist the discovery of new CAZymes and associated gene clusters in fungal genomes. Two specific aims are proposed: (1) develop new bioinformatics tools that can detect novel fungal CAZyme families without sequence similarity but with structural similarity and semantic similarity to known CAZymes, (2) conduct the first systematic study of fungal CAZyme gene clusters that have proved to be the most important mechanism for carbohydrate degradation in the bacterial world.

Project Methods

United States Department of Agriculture Project Initiation

We have two specific aims: Aim 1: develop a protein 3D structure-guided protein language model (SPLM) for fungal CAZyme prediction beyond sequence similarity. We will develop a Python software package called dbCAN-SPLM that includes SPLM model training and prediction modules. The evaluation will use two fungal CAZyme datasets that are not seen in model training as the independent benchmark datasets. Aim 2: develop a new computer program for the first systematic discovery of fungal CGCs and identify new CAZyme families from functionally unknown proteins in CGCs. We will develop CGC-Finder2 as a Python software package to define the optimized parameters for CGCs identification in all fungal genomes using a rigorous statistical simulation. New CAZymes will be discovered from unknown proteins encoded in fungal CGCs based on structure similarity and semantic similarity. The evaluation will be based on a ground truth data with 10~20 recently defined CAZyme families that include fungal members. Four new open-source software tools will be developed including the dbCAN-SPLM program, the CGC-Finder2 program, the dbCAN web server update, and the dbCAN-seq database update. Impact of these tools will be evaluated by the GitHub downloads, website access counts, user feedbacks, and paper citations.