United States Department of Agriculture
Project Initiation

Project Data for Accession Number 1034396

Project Title

SYSRV: SYSTEMS BIOLOGY REVERSE VACCINOLOGY, AN INTEGRATED VACCINE PREDICTION TOOL FOR LARGE OMICS DATA

Project Details

Sponsoring Institution National Institute of Food and Agriculture	**Project Status** ACTIVE
Funding Source NIFA Non Formula	**Grant Year** 2026
Grant No. 2026-67022-45862	**Cumulative Award Amt.** $591,500.00
Proposal No. 2024-13369	**Multistate No.** (N/A)
Project Start Date May 15, 2026	**Project End Date** May 14, 2029
Program Code [A1541] Food and Agriculture Cyberinformatics and Tools	

Project Director

Pirim, H.

Recipient Organization

NORTH DAKOTA STATE UNIVERSITY

1340 ADMINISTRATION AVE

Fargo, NORTH DAKOTA 58105

Performing Department

(N/A)

Non Technical Summary

Bacterial disease is a major threat to U.S. catfish farming, affecting fish health, farm profitability, and the stability of an important domestic food supply. A USDA report cited in this project notes a nearly 50% decline in domestic catfish production (2005-2013), with disease identified as a major contributing factor. Farmers often have limited tools including reducing feed (which slows growth and reduces production) or relying on antibiotic-medicated feeds, which can contribute to antibiotic resistance and the spread of multi-drug resistant elements. Despite substantial research effort, there are no commercially available vaccines for key bacterial pathogens affecting catfish, including Aeromonas hydrophila, Flavobacterium covae, and Edwardsiella piscicida. This gap matters beyond the farm: healthier fish means more reliable food production, stronger rural economies connected to aquaculture, and reduced public concern around antibiotic use.This project will create a practical, user-friendly web-based vaccine discovery tool that helps researchers

identify the most promising vaccine targets from large amounts of pathogen genetic information by combining proprietary and public sequence data and producing ranked candidate lists for testing. Importantly, the project will not stop at computer predictions: the team will produce and test candidates in repeated validation cycles (about 10 candidates per cycle across six cycles) and use the results to improve future selections. The successful completion of the project will speed up the path to effective vaccines, reduce losses from disease, lower reliance on antibiotics, and strengthen the resilience and sustainability of U.S. catfish aquaculture.

Animal Health Component

Animal Health Component: 100%

Research Effort Categories

Basic	0%
Applied	100%
Developmental	0%

Classification

Knowledge Area (KA)	Subject of Investigation (SOI)	Field of Science (FOS)	Percent
135	4010	2080	70%
315	3710	1040	30%

Knowledge Area

[135] Aquatic and Terrestrial Wildlife

[315] Animal Welfare/Well-Being and Protection

Subject Of Investigation

[4010] Bacteria

[3710] Catfish

Field Of Science

[2080] Mathematics and computer sciences

[1040] Molecular biology

Keywords

Reverse vaccinology Vaccine design Network analysis Machine learning Deep learning Optimization

Goals / Objectives

Goal: Develop and deliver A-SysRV, an integrated systems-biology reverse vaccinology platform for U.S. catfish aquaculture. This platform combines proprietary and public pathogen omics/genomics data with protein language model embeddings, graph neural networks for predicting hypothetical protein functions, machine-learning-based vaccine candidacy scoring, and optimization-based selection. Iteratively, A-SysRV improves prediction quality through

in vivo validation, enabling researchers to efficiently identify high-confidence vaccine candidates and accelerate the development of effective vaccines for key bacterial pathogens impacting catfish production.Objectives Organize core genome and protein sets for target aquaculture pathogens, such as Aeromonas, Edwardsiella, and Flavobacterium species, and curate pathogen datasets. Extract the associated hypothetical proteins required for downstream modeling. Construct protein relationship networks, including similarity and interaction graphs, using database-derived features and ESM-2 sequence embeddings. These networks should also include hypothetical proteins and be ready for network analysis. Predict hypothetical protein functions by applying GNN-based learning on the constructed networks, including inference of missing relationships and storing predicted functional annotations for hypothetical proteins. Implement vaccine-relevant filtering for proteins (including hypothetical proteins) by computing and organizing key attributes used in reverse vaccinology--e.g., subcellular localization, adhesin probability, transmembrane domains, immunogenicity, and physicochemical properties--into a consistent feature set for modeling. Train and validate vaccine-candidacy prediction models (ML/DL classifiers) using protective-antigen resources and curated literature datasets, selecting the best-performing model(s) via standard evaluation metrics and producing probability/ranking outputs for candidate selection. Select top candidates via optimization by deploying an optimization-based selection module (e.g., mixed-integer programming) that chooses the best k vaccine candidates per pathogen using model scores and user-defined priorities over vaccine-related features, returning an interpretable ranked set for testing. Biologically validate predictions in vivo and close the loop by conducting iterative channel catfish vaccination /challenge and immune-response assays, validating approximately 10 candidates per cycle across multiple cycles (targeting ~60 candidates total), and using outcomes to update/improve the predictive models. Deploy a scalable web-based tool for end users by delivering a reliable web interface that provides access to data, analyses, and vaccine prediction outputs (desktop/mobile compatible), enabling researchers to use A-SysRV efficiently and consistently.

Project Methods

The project will be conducted as an end-to-end reverse vaccinology (RV) discovery + validation + deployment pipeline that integrates computational prediction with iterative wet-lab confirmation and a web-delivered decision-support tool. On the computational side, we will curate pathogen core genomes/proteomes and derive biological/physicochemical features, then construct protein networks and apply graph-based learning to predict missing relations and hypothetical protein (HP) functions as part of the RV prioritization workflow. For vaccine-candidacy prediction, we will train multiple ML/DL models using Protegen and literature-derived protective antigen data, with explicit data splitting into training /validation/test sets, parameter tuning on validation data, and best-model selection on test data. Model performance will be analyzed using precision, recall, and F1 to address the class imbalance inherent in identifying protective antigens. Data-quality steps include preprocessing to handle missing values, standardize features, and align multi-source inputs, with robustness supported via cross-validation and grid-search style hyperparameter tuning.On the experimental validation side, candidate validation is designed as an integral component of tool development. We will express and purify selected antigens and evaluate candidates in repeated validation cycles; each cycle evaluates ~10 candidates, and there are six validation cycles (? 60 candidates total). Validation results (e.g., protective efficacy /immune response outcomes) will be statistically analyzed and then used to update the ML models, creating a feedback loop that improves the computational prioritization over time.The delivery mechanism for reaching the target audience is a web application that integrates filtering, ML, and optimization modules in a single interface. A unique implementation aspect is the planned microservice architecture for the web application (with a stated contingency plan for technology-stack challenges), chosen to support maintainability and sustainability. Sustainability and reach are further supported by open-source languages/tools and an explicit plan to open-source SysRV and host it on GitHub, inviting external contributors.Efforts that deliver science-based knowledge to people will include: Practicum-style research training and mentoring of four graduate students (computational pipeline, ML/selection, software/UI /database, and wet-lab validation roles).User-oriented dissemination of the tool and results through conference presentations and publications, plus discoverability activities (SEO) so users can find and directly access the tool's results.Open-source distribution and community engagement via GitHub hosting to enable reuse/extension and broaden adoption.Evaluation will connect project milestones to measurable indicators: Computational model evaluation (change in knowledge/action: better prioritization): Indicators: precision/recall/F1 on held-out test data for

United States Department of Agriculture
Project Initiation

vaccine-candidacy models. Data collected: training/validation/test performance logs; hyperparameter search results; versioned datasets and models used for each evaluation cycle. Success criterion: demonstrated model performance using the stated metrics, and improved/updated models after incorporating validation-cycle outcomes.Wet-lab validation evaluation: Indicators: number of candidates expressed/purified and tested; number of validation cycles completed; number of candidates evaluatedand number showing "successful" vaccine performance per cycle. Data collected: assay results (challenge/immune response), protocol logs, and statistical analysis outputs used to interpret efficacy.Project monitoring and milestone evaluation: Indicators: completion of task-level milestones across years (including start of model-deployment testing and maintenance activities) and on-time delivery of annual reporting. Process: monthly (or more frequent) team meetings with documented minutes to monitor progress and ensure timely completion.