Advanced Bioinformatics Project Proposal

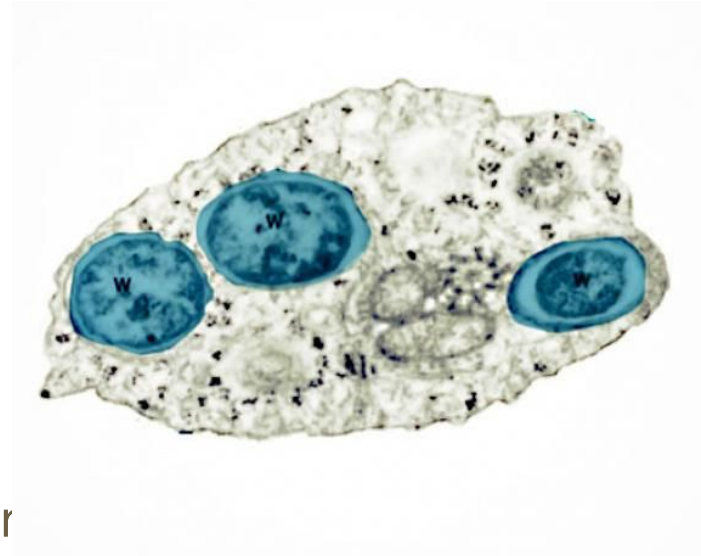
Jesse Valliere
Spring 2022

Project 1 - *Wolbachia* cif Gene Reference Database

In collaboration with Sarah Bordenstein, M.S. and the Bordenstein Lab at Vanderbilt University

Wolbachia - What are they?

- Wolbachia genus
  - Very common
  - Known to infect arthropod species
  - Live inside cells and pass on to future generations through eggs
  - When carried by mosquitoes, they have reduced ability to spread viruses to humar
    - Dengue, Zika, Chikungunya and Yellow Fever

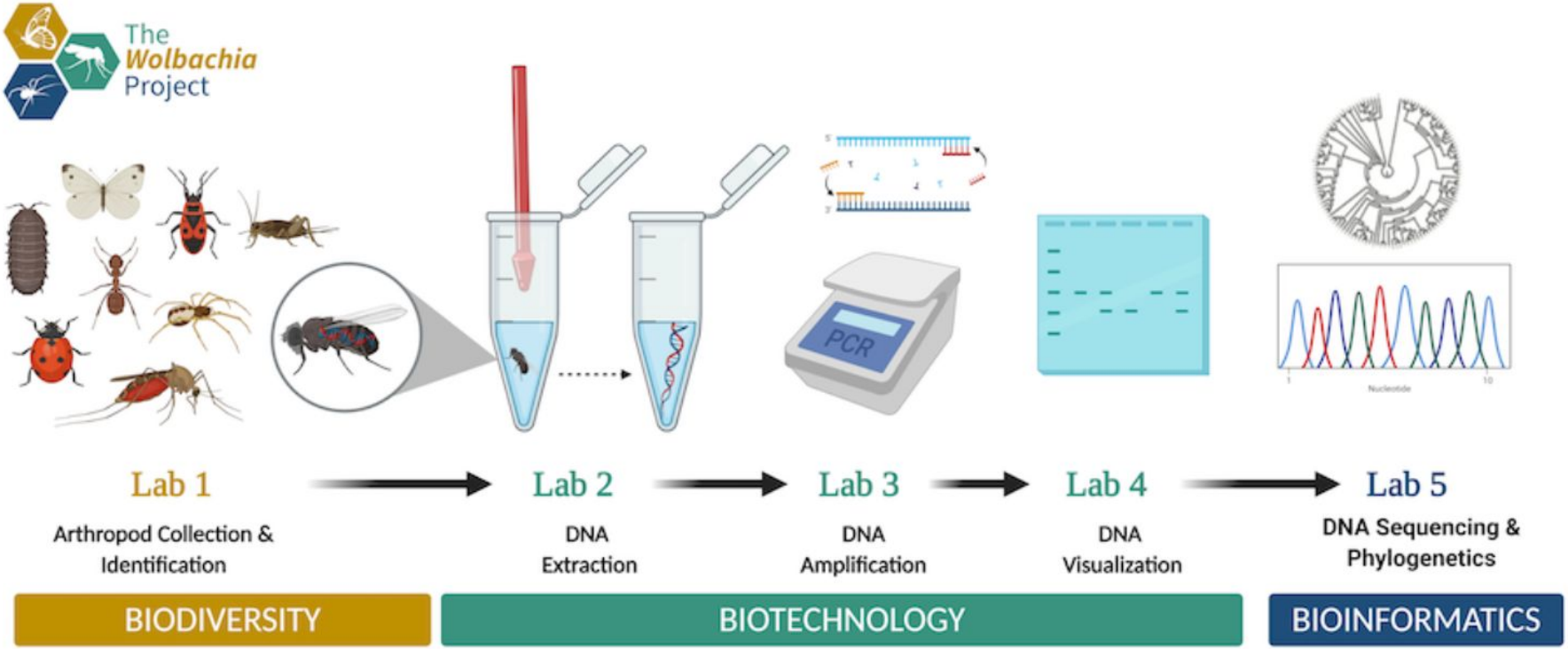


The Wolbachia Project

The Wolbachia Project

Discover the Microbes Within!

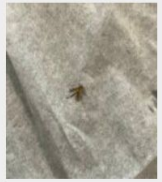
- Done in collaboration with High School educators and students
  - Obtain wild samples
  - Analyze DNA sequences of potentially new strains of Wolbachia
  - Upload to repository for users to recognize and use
- Have the ability to collaborate globally, as it can be done from practically anywhere


The Wolbachia Project
PCR
Lab 1
Arthropod Collection & Identification
Lab 2
DNA Extraction
Lab 3
DNA Amplification
Lab 4
DNA Visualization
Lab 5
DNA Sequencing & Phylogenetics
BIODIVERSITY
BIOTECHNOLOGY
BIOINFORMATICS

Example entry in The *Wolbachia* Project Database

Also includes methods performed, Buffers used, results, and confidence levels

Sample information

Picture	 Photos by: Stanley T.
Location	
Collection date	09/13/2022
Captive / Cultivated?	Wild-caught
Group	Acton-Boxborough Regional High School
Observations	

Database Ideas

- While The *Wolbachia* Project Database includes a lot of valuable information in regards to *Wolbachia* in wild species, it is only surface level.
  - We know if they do or do not have *Wolbachia*, but in order to make more use of this data we need to know more about the genes these species are expressing
- For example, if they are expressing cifA;cifB genes, these mosquitoes have a decreased risk in spreading specific arboviruses.

What the Database would have…

- After speaking to Sarah Bordenstein, the director and co-founder of The W*olbachia* Project, the main ideas for the database would be centered around the cifA and cifB genes and would include:
  - Reference entries for each cifA or cifB sequence, including Wolbachia strain; arthropod host; location; Cif type; etc
  - Some kind of upload feature so that people could add their own sequences as they are discovered
  - Incorporation of a BLAST tool that would check against the reference database and help determine if the sequence is a Cif.

Possible limitations

- Over 100 different cifA and cifB gene sequences already determined
  - Slightly smaller data set, but it would expand over time as more are determined
- Database maintenance Post-Graduation
  - How can I keep it up to date and properly working after I graduate without continuous attention?
  - Is there a way I could autonomously check uploaded sequences and ensure it accurately incorporates new ones?
- Special thank you to the Bordenstein Lab and Sarah Bordenstein!!

The Wolbachia Project

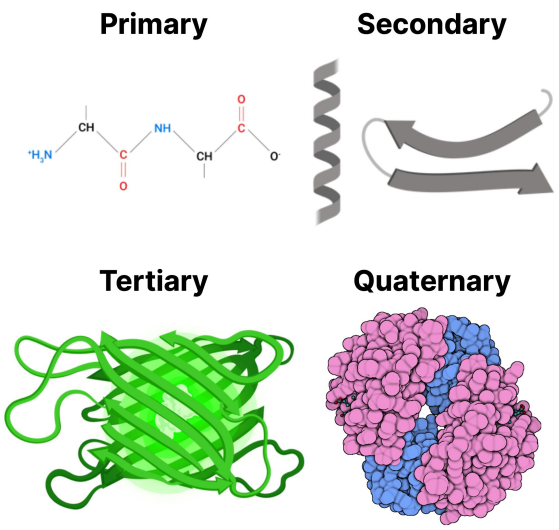
Discover the Microbes Within!

Project 2 - Amino Acid Sequence Analysis Tool

To examine important biological properties such as domains, motifs, signal sequences, hydrophobicity, structure, etc. using multiple different tools

Bioinformatic Tools & Software

- Understanding amino acid properties can be critical to protein structure and function determination.
- Example properties such as Hydrophobicity, Secondary Structure, Motifs, Domains, etc. can allow researchers to accurately determine this.
- While it used to be found mainly through wet-lab approaches, can be completely automated in a computational format for practically no cost
    - Results can usually be statistically significant and accurate, and can be confirmed with wet-lab techniques



Bioinformatics Project

- Utilization of software tools to perform an analysis on an unknown sequence
  - Separated into categories: Functional/Structural Domains & Motifs, Sequence Motifs, Secondary Structure & Hydrophobicity, Protein & Domain interactions, Transmembrane domains, subcellular localization, 3D structure
- While it was great to do, felt time consuming for something that could be made autonomous
  - Had to go to each website, input sequence, save results, then compare them, etc.

What This Tool would do

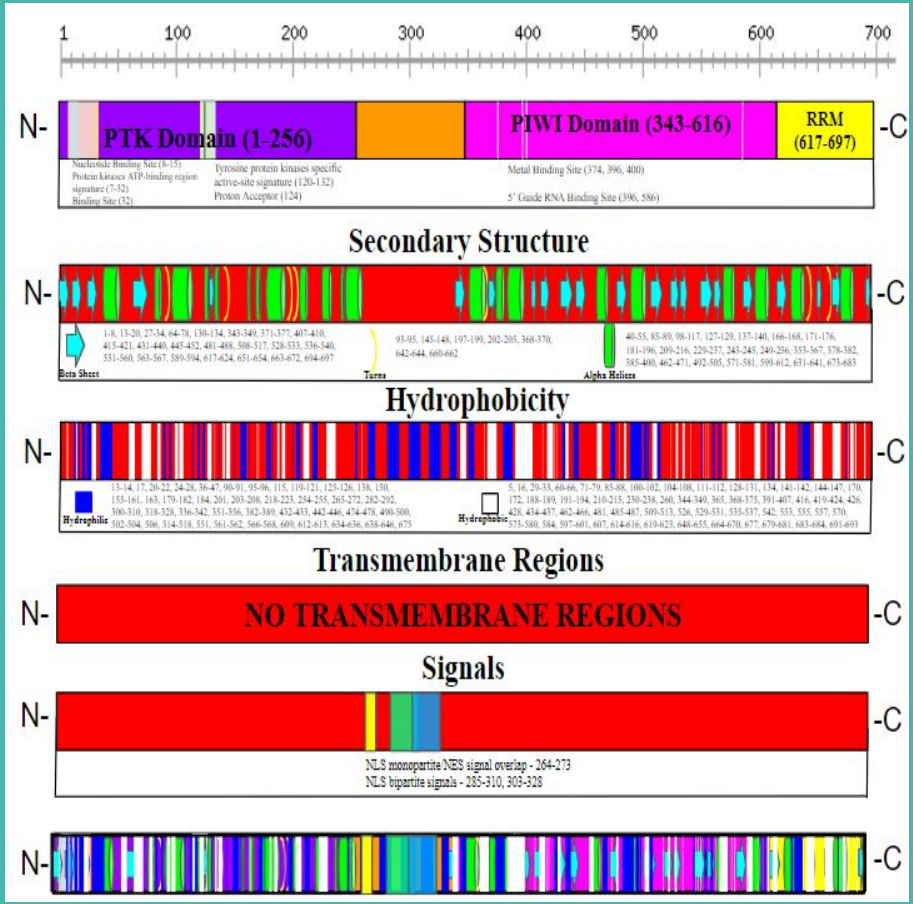
- Allow user to input sequence of interest
- Check if this is a known, characterized protein sequence.
    - If it is, then it can save time and just return the already known data
- Otherwise, begin sending sequences to respective websites through back-end
- Compile results, and check for similarities among different tools
    - Overlap results to see how similar
- Create a interactive sequence tool to highlight domains, motifs, secondary structures, etc. in the sequence itself.
- Save runs for certain amount of time so they can be reviewed

Example of sequence comparison

Would be compiled from all results and include more than this UniProt example!

Example of output format

Would be able to choose which results (from a menu) you want to be used to create this overlay:



Possible Limitations

- Retrieving Data could be very consuming and slow
  - I would want it to be practical and use as much offline computing as possible to save time
    - Ex: hydrophobicity I would prefer to use an algorithm over a website since results would be returned faster
- Data can be classified differently in tools but may mean the same thing
  - Not exactly sure how I could compare them or if I should bother
  - Ex: PIWI Domain result in tool A may be PIWI_Cyto Domain in tool B

Thank you!