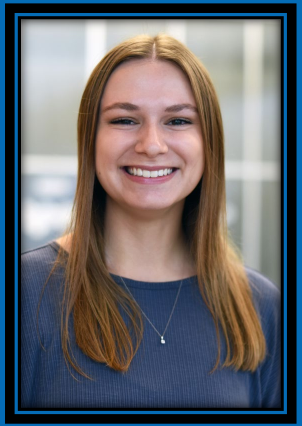
United States Environmental Protection Agency

Sequence Alignment to Predict Across Species Susceptibility (SeqAPASS) v7.1

Instructors:

Carlie A. LaLone, Ph.D. (US EPA)

Marissa Brickley, PhD student (UMN-Duluth/ US EPA)



Office of Research and Development
Center for Computational Toxicology and Exposure, Great Lakes Toxicology and Ecology Division

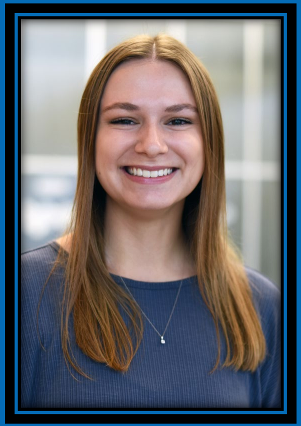
United States
Environmental Protection
Agency

Introduction - SeqAPASS v7.1:

Instructors:

Carlie A. LaLone, Ph.D. (US EPA)

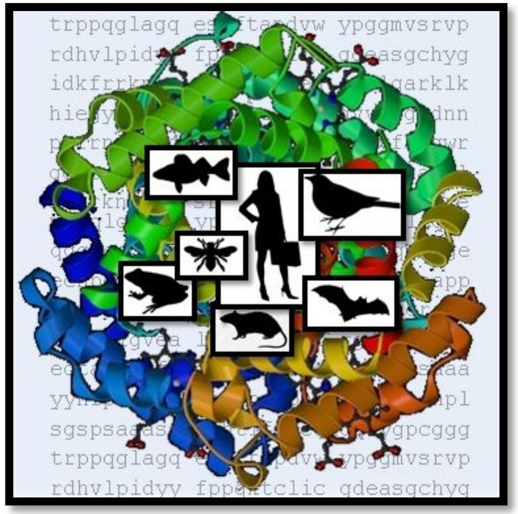
Marissa Brickley, PhD student (UMN-Duluth/ US EPA)

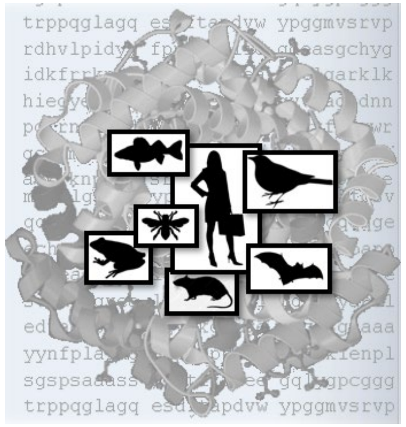


Office of Research and Development
Center for Computational Toxicology and Exposure, Great Lakes Toxicology and Ecology Division

Training Course Overview

- **Introduction** to challenges in species extrapolation and how bioinformatics can aid in addressing these challenges (LaLone)
- Training with SeqAPASS
  - Level 1 and Level 1 Data visualization (LaLone)
  - Level 2 and Level 2 Data visualization (Jensen)
    - Instructor Demos – Level 1 and 2
  - Level 3 and Level 3 Data visualization (Jensen)
    - Instructor Demos – Level 3
  - Decision summary report (Jensen)
  - Level 4 (LaLone)

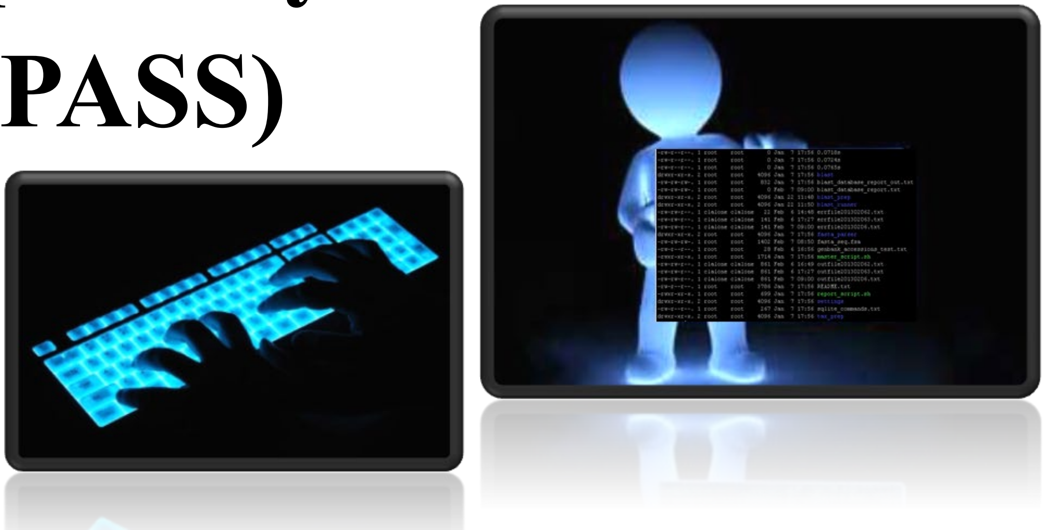




https://seqapass.epa.gov/seqapass/

Sequence Alignment to Predict Across Species Susceptibility

(SeqAPASS)



OXFORD

SOT | Society of Toxicology

www.toxsci.oxfordjournals.org

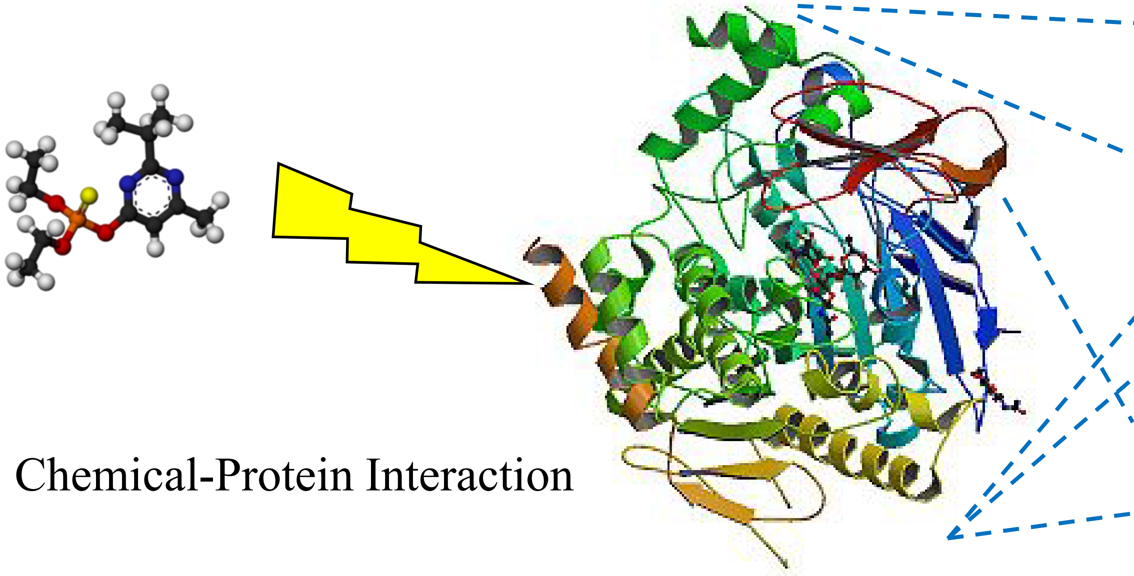
TOXICOLOGICAL SCIENCES, 153(2), 2016, 228–245

doi: 10.1093/toxsci/kfw119
Advance Access Publication Date: June 30, 2016
Research article

Sequence Alignment to Predict Across Species Susceptibility (SeqAPASS): A Web-Based Tool for Addressing the Challenges of Cross-Species Extrapolation of Chemical Toxicity

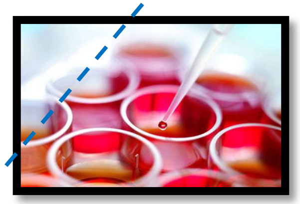
Carlie A. LaLone,[*,1] Daniel L. Villeneuve,[*] David Lyons,[†] Henry W. Helgen,[‡]
Serina L. Robinson,[§,2] Joseph A. Swintek,[¶] Travis W. Saari,[*] and
Gerald T. Ankley[*]

What information is required for a SeqAPASS query?



Chemical-Protein Interaction

Knowledge of a sensitive or targeted species



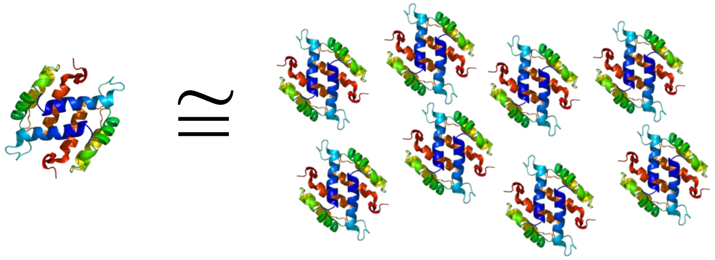
Knowledge of the model organism used in an *in vitro* assay

1. Protein
2. Species

MIE → KE → KE → KE → AO

Knowledge of the species for which the Key Event was developed

Chemical Molecular Target
in Target Species



Compare to Millions of Proteins

From Thousands of Species

Greater similarity = Greater likelihood that chemical can act on the protein

Line of Evidence: Predict Potential Chemical Susceptibility Across Species

Gather Lines of Evidence Toward Protein Conservation

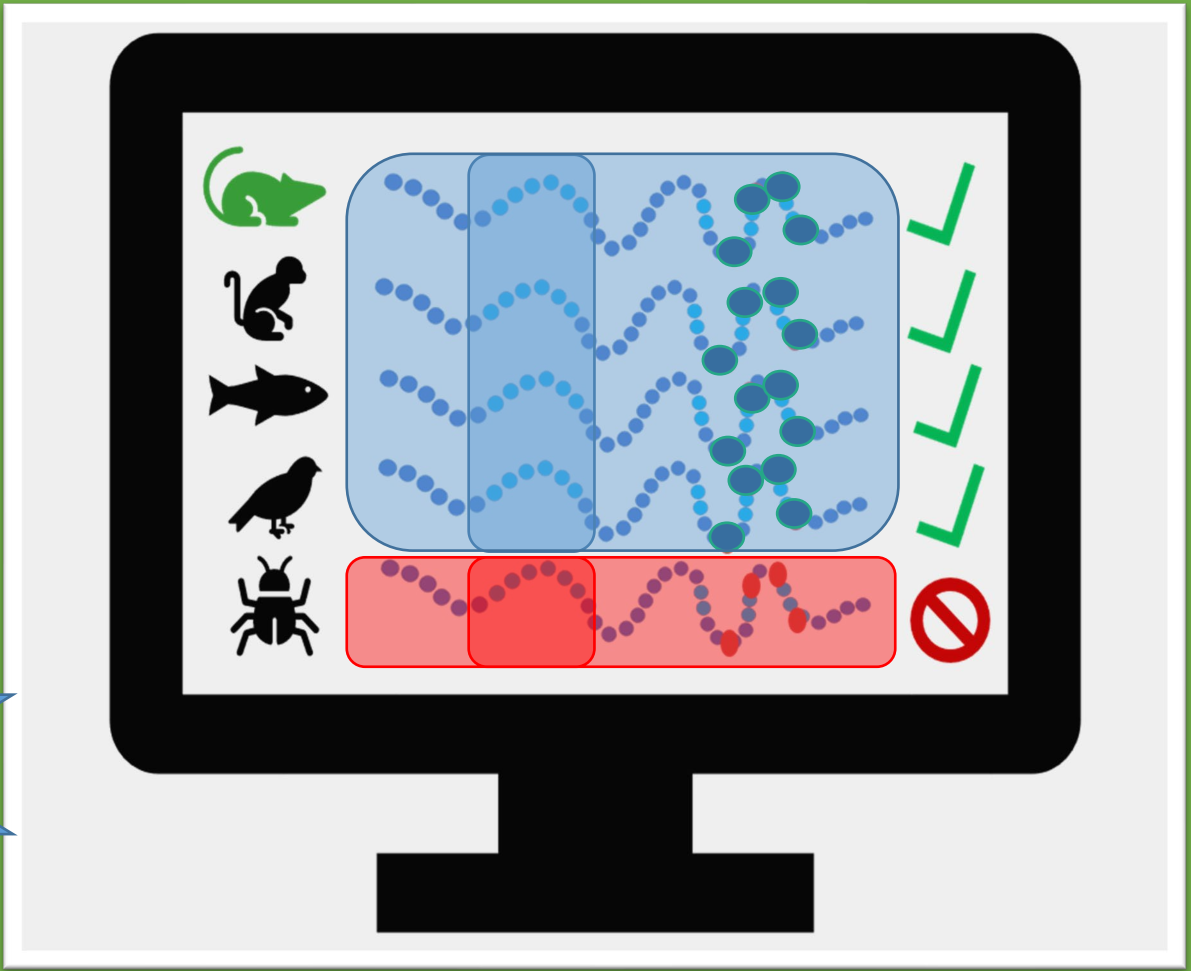
Flexible Analysis Based On Available Data

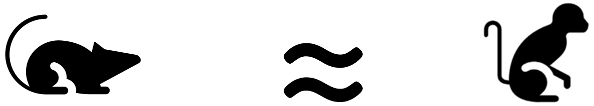
Level 1 Primary Amino Acid Sequence Alignments

Level 2 Conserved Functional Domain Alignments

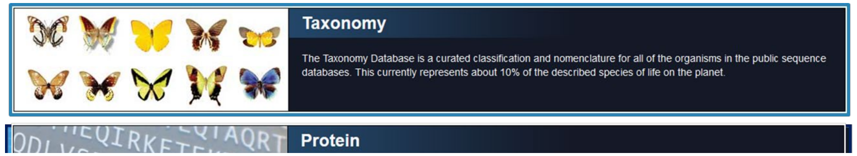
Level 3 Critical (Close Contact) Amino Acid Conservation

seqapass.epa.gov/seqapass/

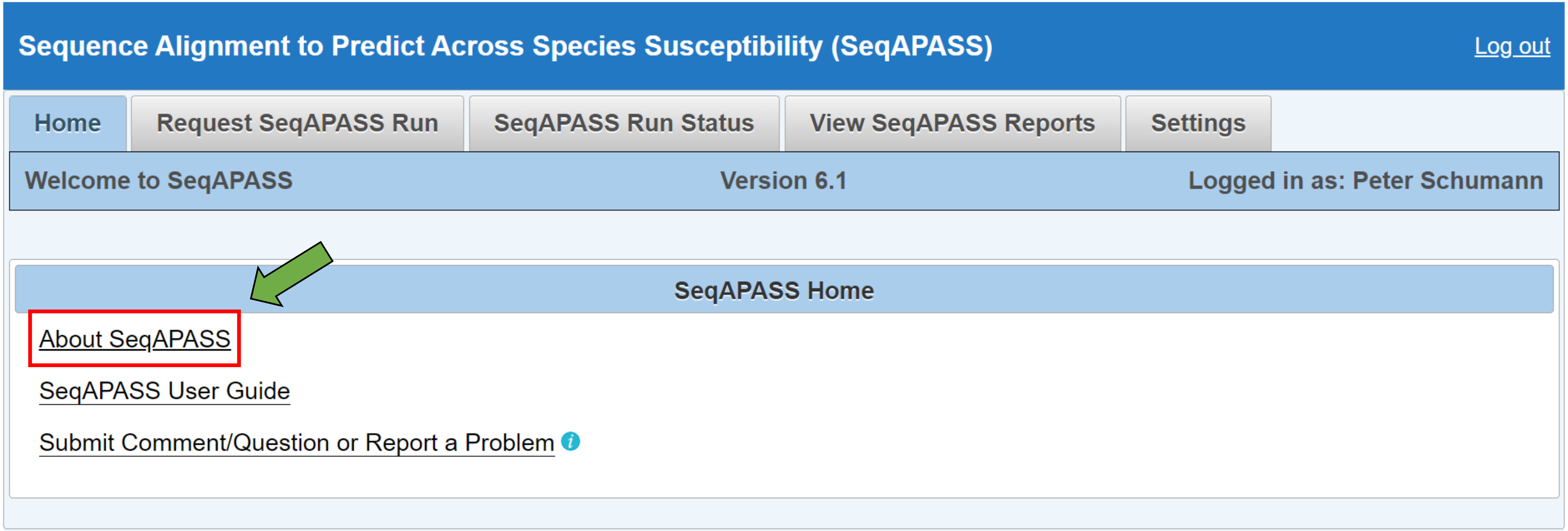




SeqAPASS uses publicly available data and tools:



The Taxonomy Database is a curated classification and nomenclature for all of the organisms in the public sequence databases. This currently represents about 10% of the described species of life on the planet.



Sequence Alignment to Predict Across Species Susceptibility (SeqAPASS)

Welcome to SeqAPASS

Version 6.1

Logged in as: Peter Schumann

SeqAPASS Home

About SeqAPASS

SeqAPASS User Guide

Submit Comment/Question or Report a Problem

4	05/15/2019 10:45 AM	3.2	2.8.1	02/28/2019	12/08/2016	07/09/2010	0.75
3	03/04/2019 05:54 PM	3.1	2.6.0	10/25/2017	12/08/2016	07/09/2010	0.75
3	03/10/2018 02:12 AM	3.0	2.6.0	10/25/2017	12/08/2016	07/09/2010	0.75
2	05/24/2017 06:59 PM	2.0	2.5.0	01/04/2017	02/05/2016	07/09/2010	0.75
1	01/27/2016 08:00 PM	1.0	2.3.0	11/09/2016	04/25/2015	07/09/2010	0.75

Threatened/Endangered Species Data obtained from EPA ECOTOX on Feb. 2, 2022

Animation by: Miguel Moravec (EPA CSS) & Andrew Patterson

SeqAPASS

Taxonomy

The Taxonomy Database is a curated classification and nomenclature for all of the organisms in the public sequence databases. This currently represents about 10% of the described species of life on the planet.

Protein

The Protein database is a collection of sequences from several sources, including translations from annotated coding regions in GenBank, RefSeq and TPA, as well as records from SwissProt, PIR, PRF, and PDB. Protein sequences are the fundamental determinants of biological structure and function.

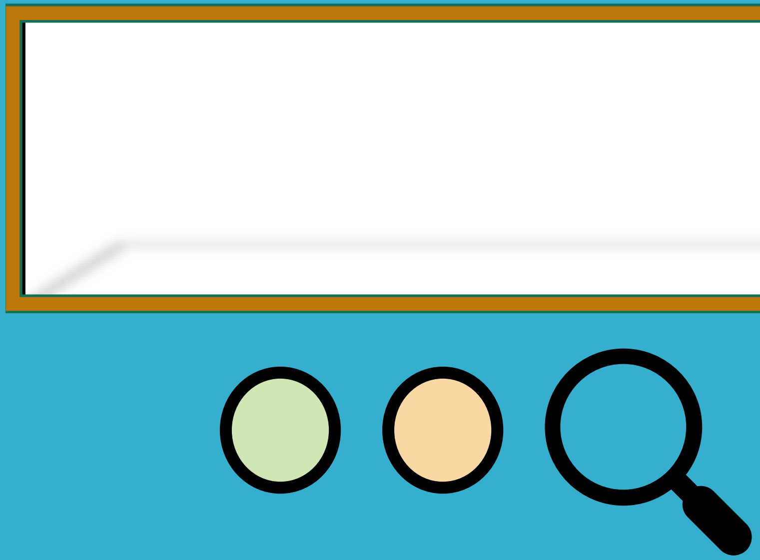
CDD

The Conserved Domain Database is a resource for the annotation of functional units in proteins. Its collection of domain models includes a set curated by NCBI, which utilizes 3D structure to provide insights into sequence/structure /function relationships.

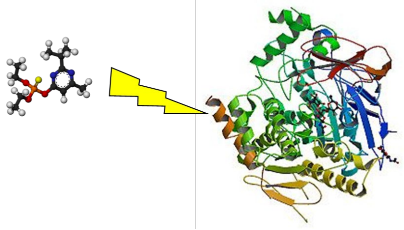
SeqAPASS

United States
Environmental Protection
Agency

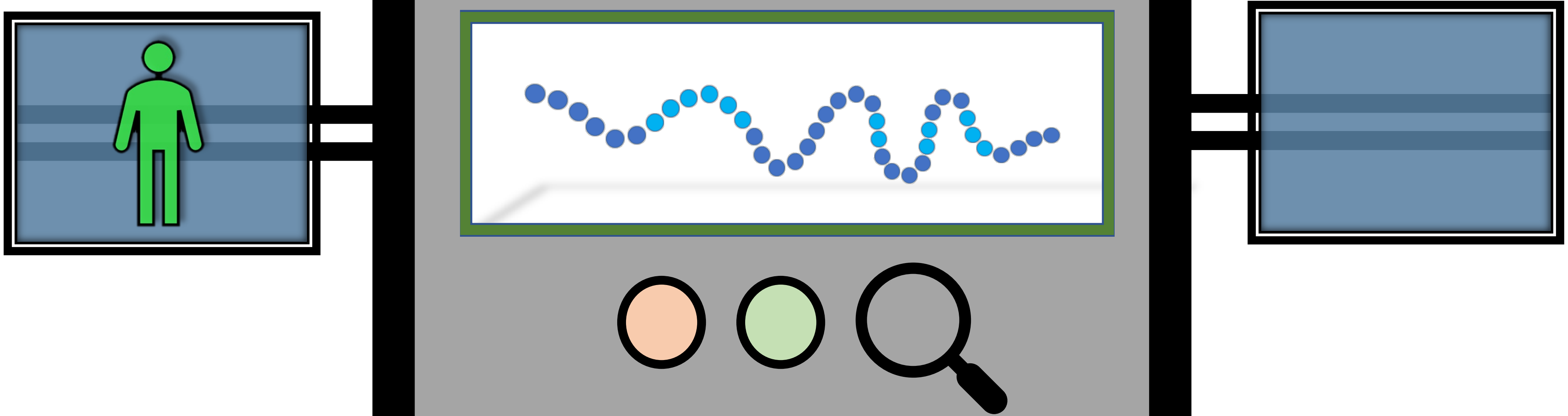
SeqAPASS
Level 1



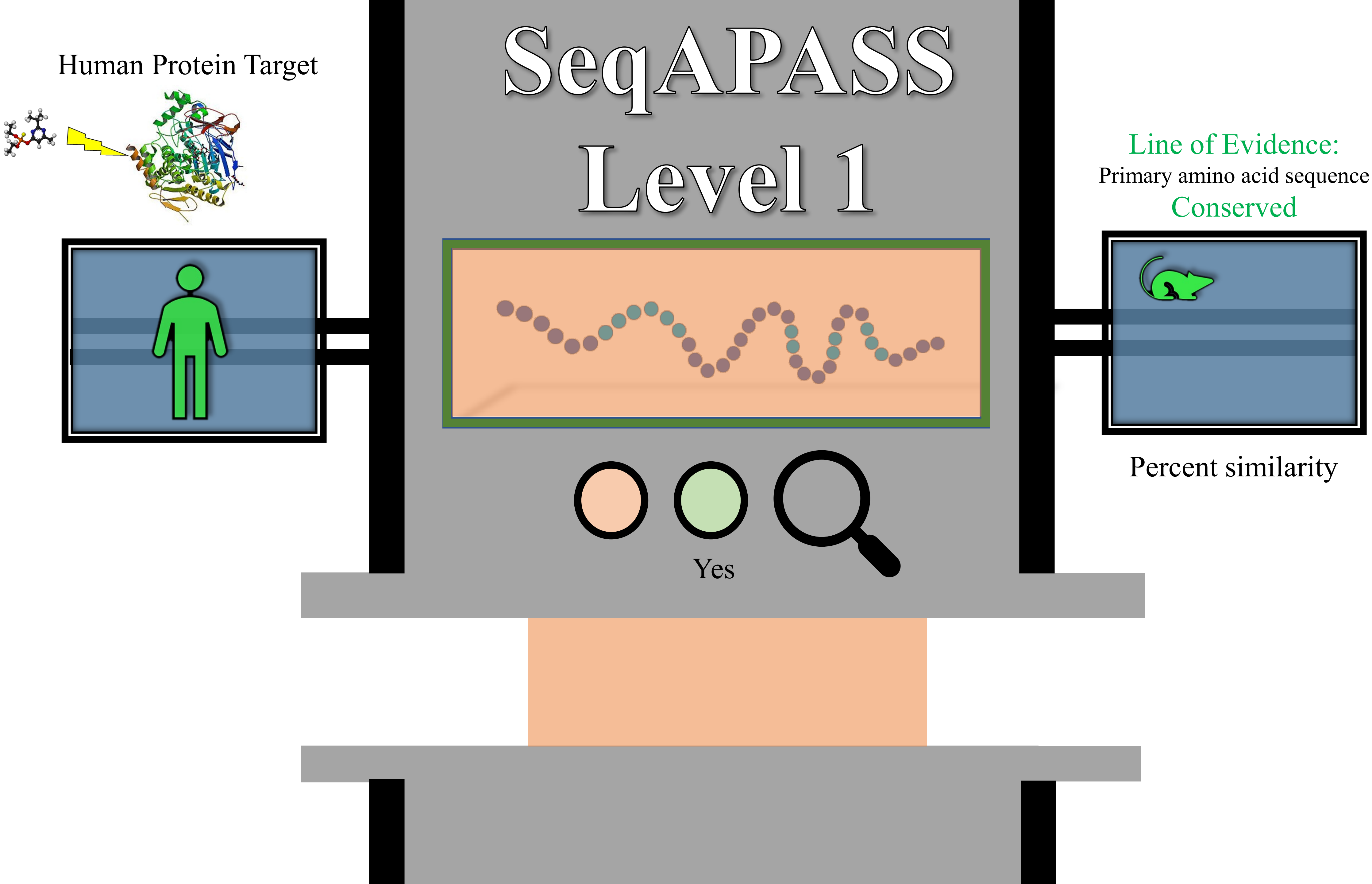
Human Protein Target



SeqAPASS
Level 1



SeqAPASS Level 1



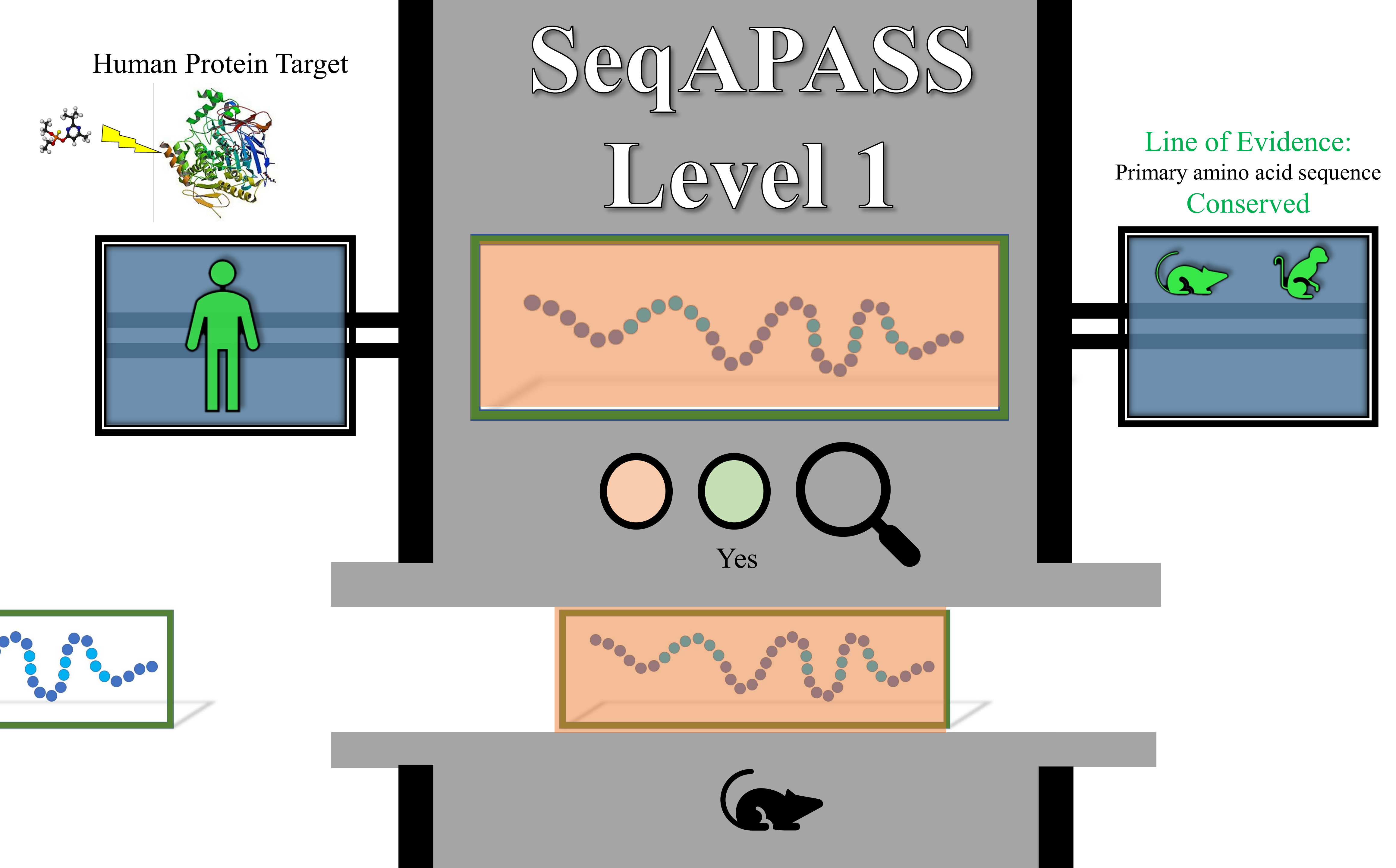
Human Protein Target

Line of Evidence:
Primary amino acid sequence
Conserved

Percent similarity

Yes

SeqAPASS Level 1



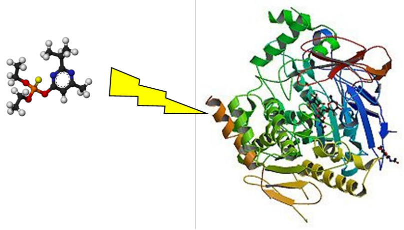
Human Protein Target

Line of Evidence:
Primary amino acid sequence
Conserved

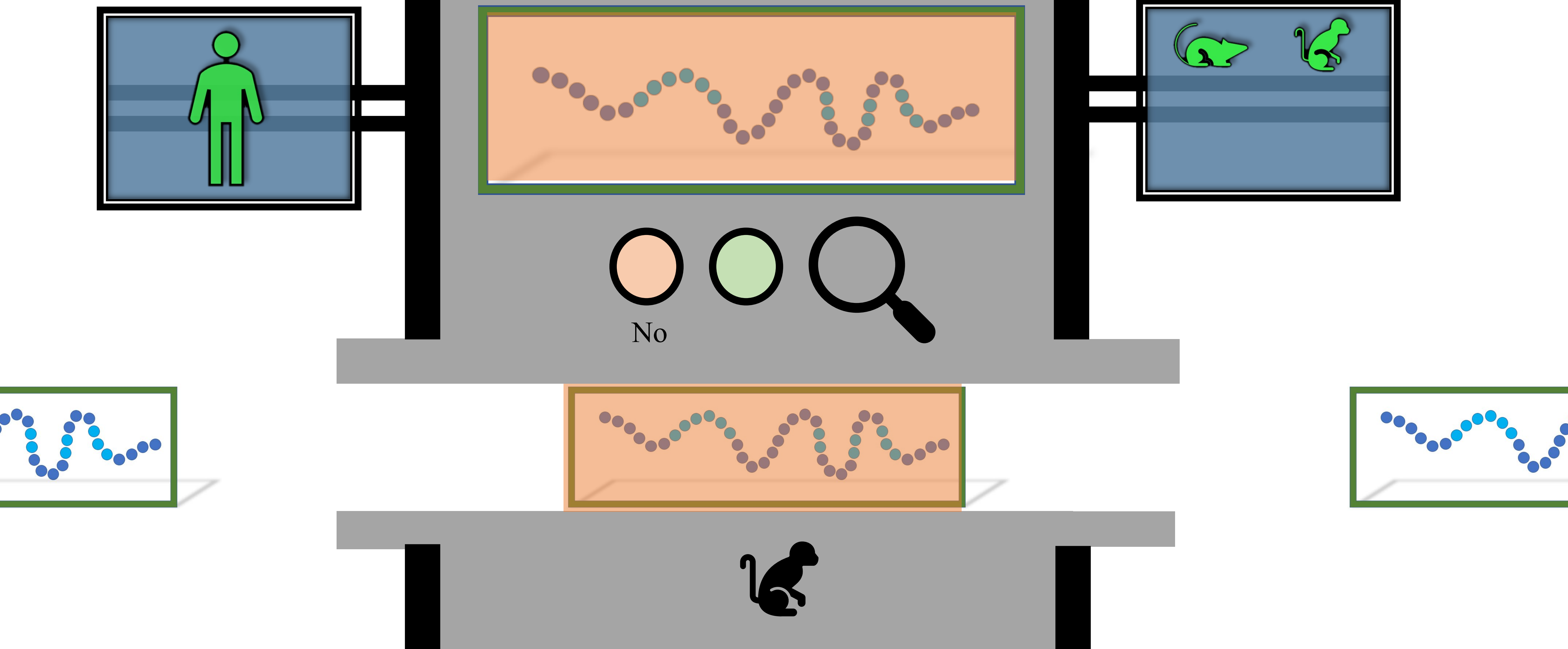
Yes

SeqAPASS
Level 1

Human Protein Target



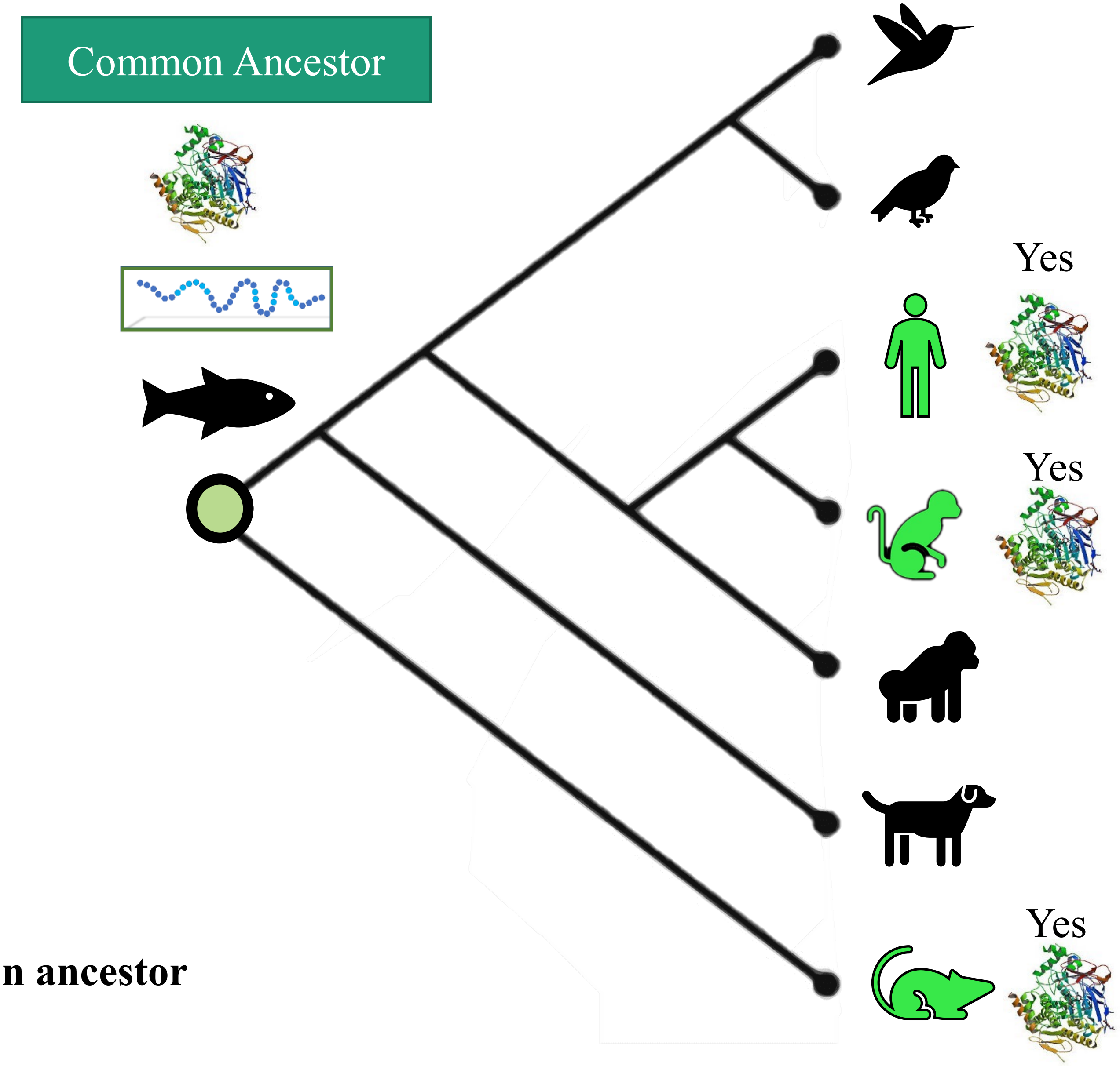
Line of Evidence:
Primary amino acid sequence
Conserved



SeqAPASS Level 1

Ortholog Candidate Identification

Yes



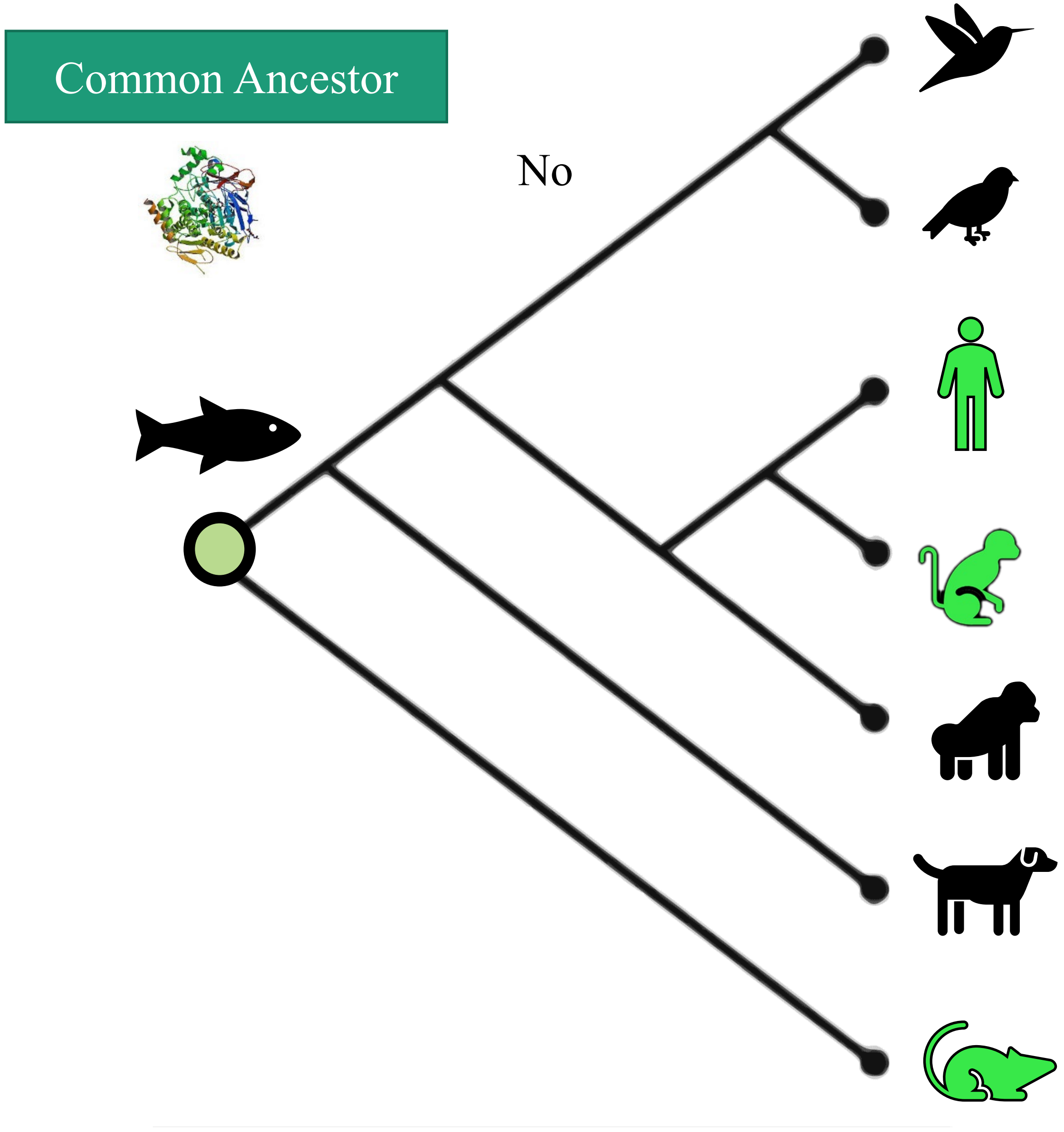
Proteins in different species that evolved from a common ancestor

Typically maintain similar function

SeqAPASS Level 1

Ortholog Candidate Identification

No



SeqAPASS Level 1

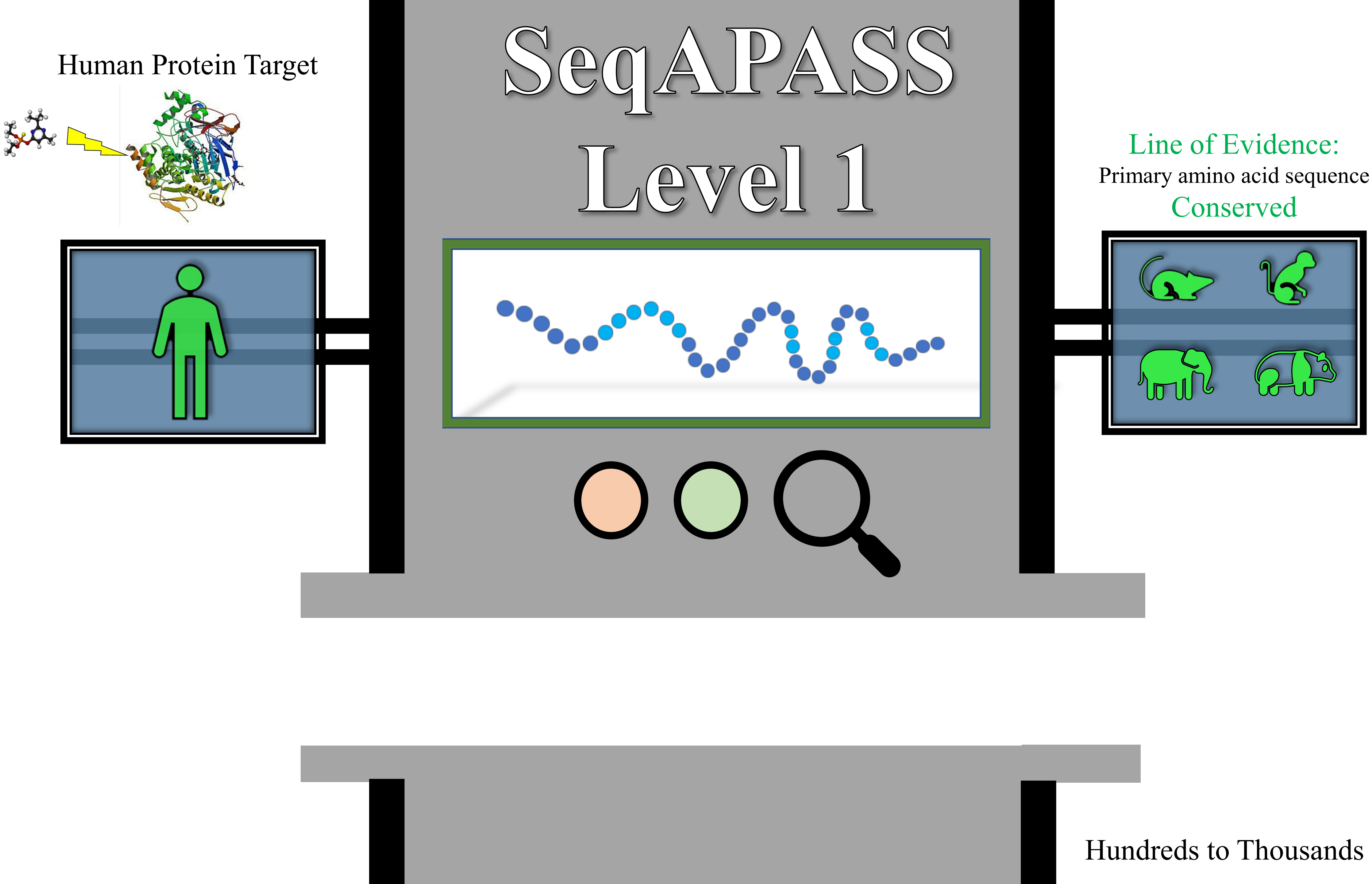
Common Name	Ortholog Candidate	Cut-off	Percent Similarity
Human	Y	33.15	100
Florida manatee	Y	33.15	98.8
Mallard	Y	33.15	82.29
Rock pigeon	Y	33.15	80.93
Green anole	Y	33.15	80.65
Pacific transparent sea squirt	Y	33.15	33.15
Yesso scallop	N	33.15	32.87
Purple sea urchin	N	33.15	26.05
Human whipworm	N	33.15	23.53
Bed bug	N	33.15	21.62

Lowest % Similarity that is still an ortholog

Example:
Susceptibility Cut-off: Set at 33.15
Above cut-off: More likely to be susceptible base on similar **FUNCTION**

SeqAPASS Level 1

Human Protein Target

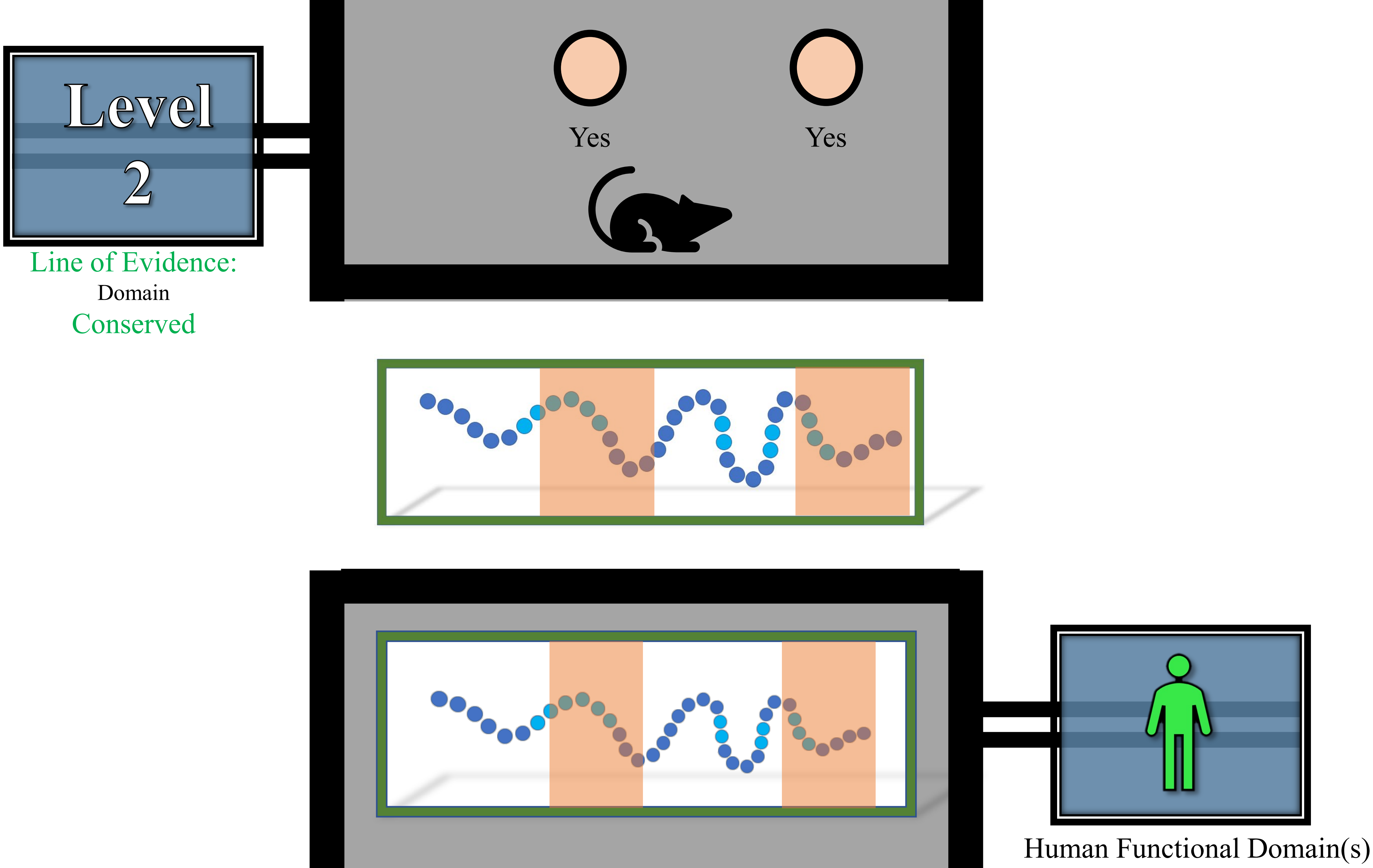


Line of Evidence:

Primary amino acid sequence

Conserved

Hundreds to Thousands of Species



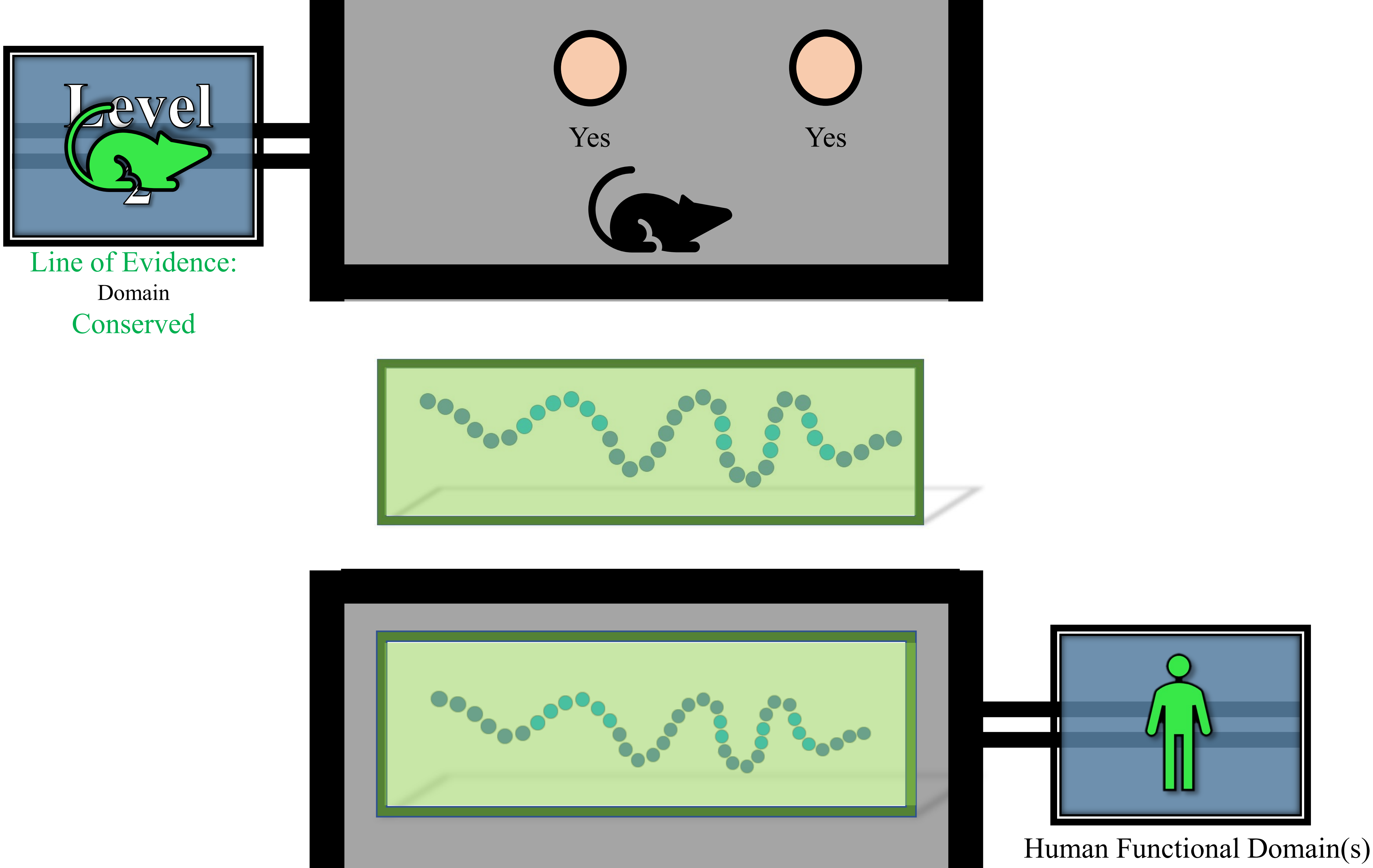
Level 2

Line of Evidence:

Domain

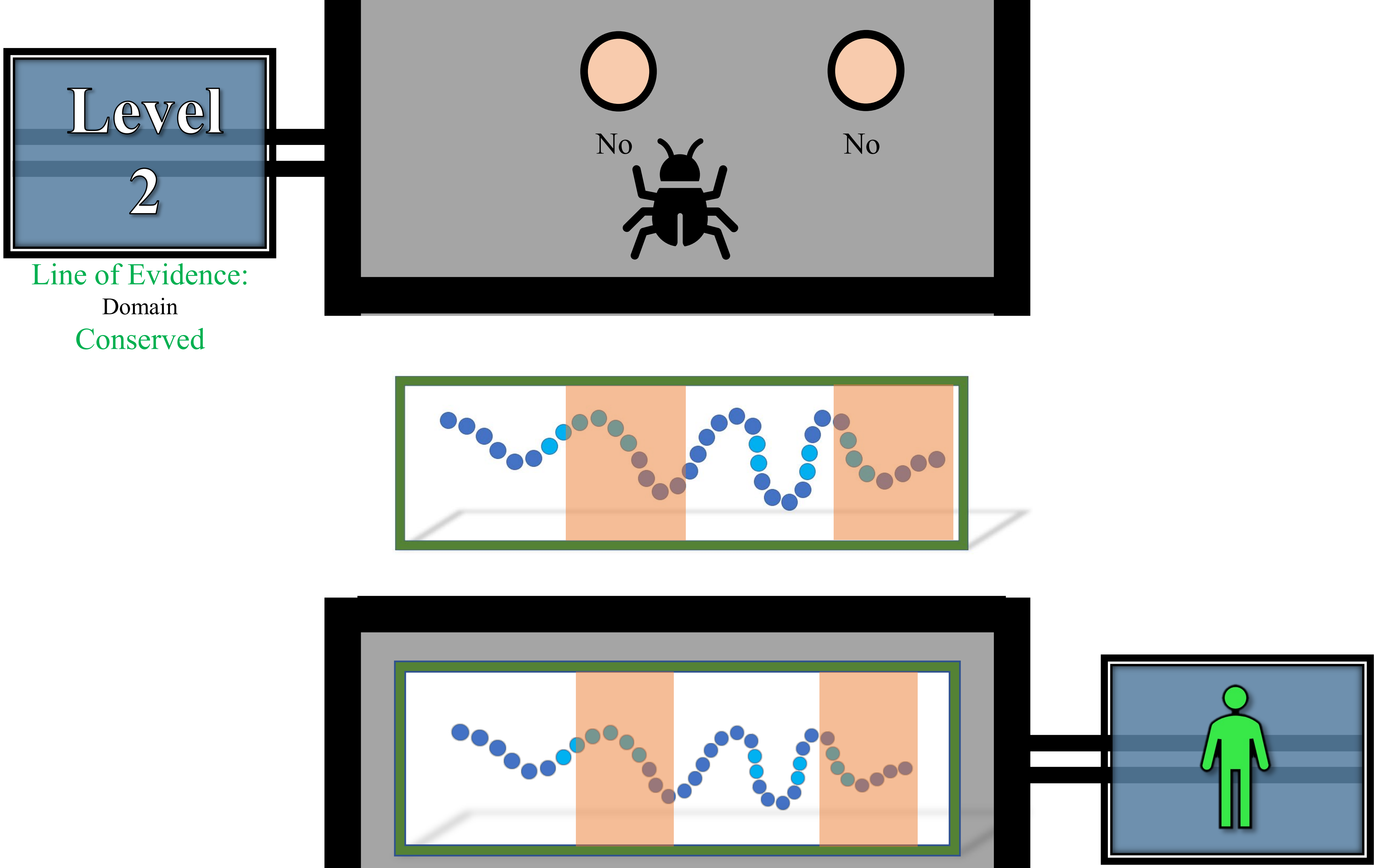
Conserved

Human Functional Domain(s)



Line of Evidence:
Domain
Conserved

Human Functional Domain(s)



Level 2

Line of Evidence:
Domain
Conserved

No

No

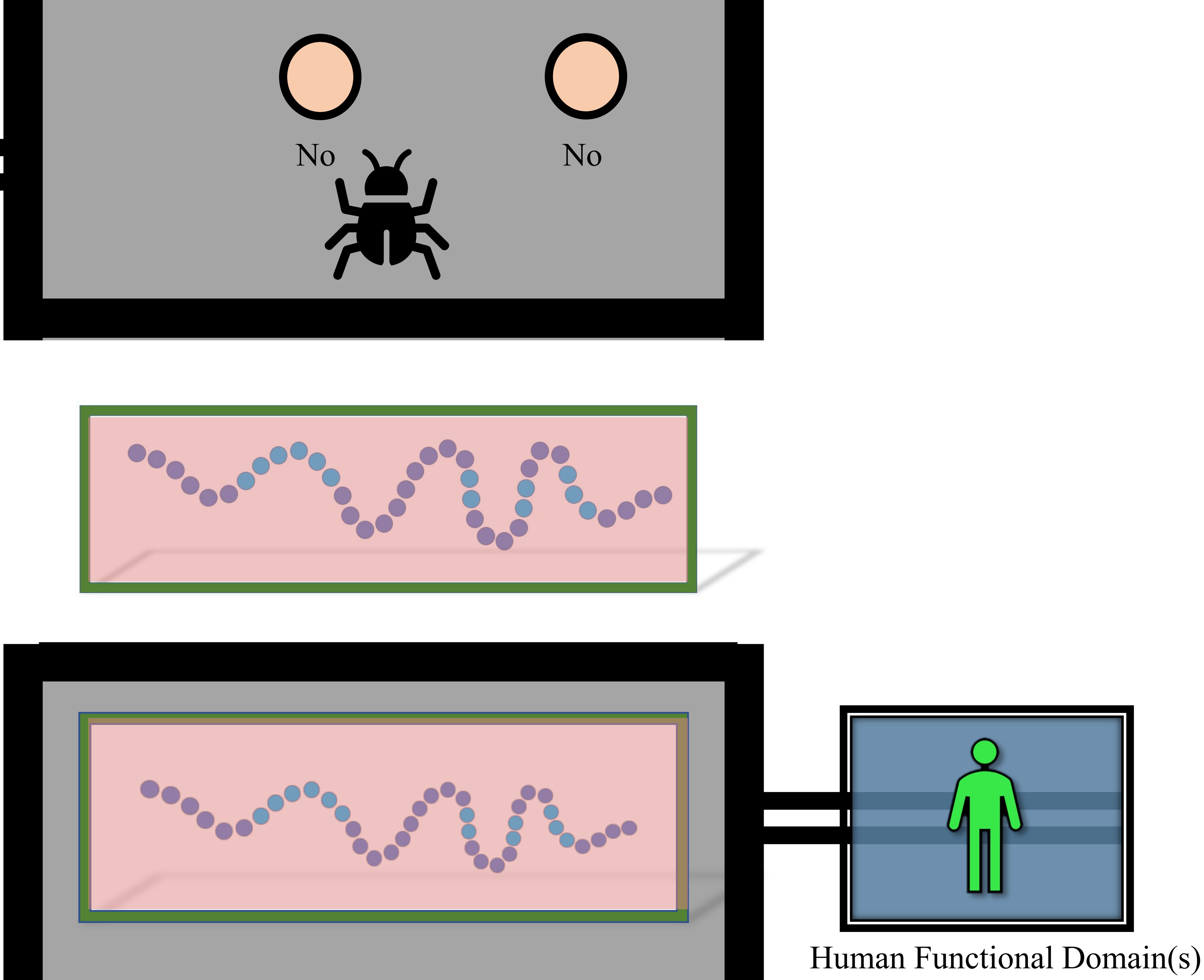
Human Functional Domain(s)

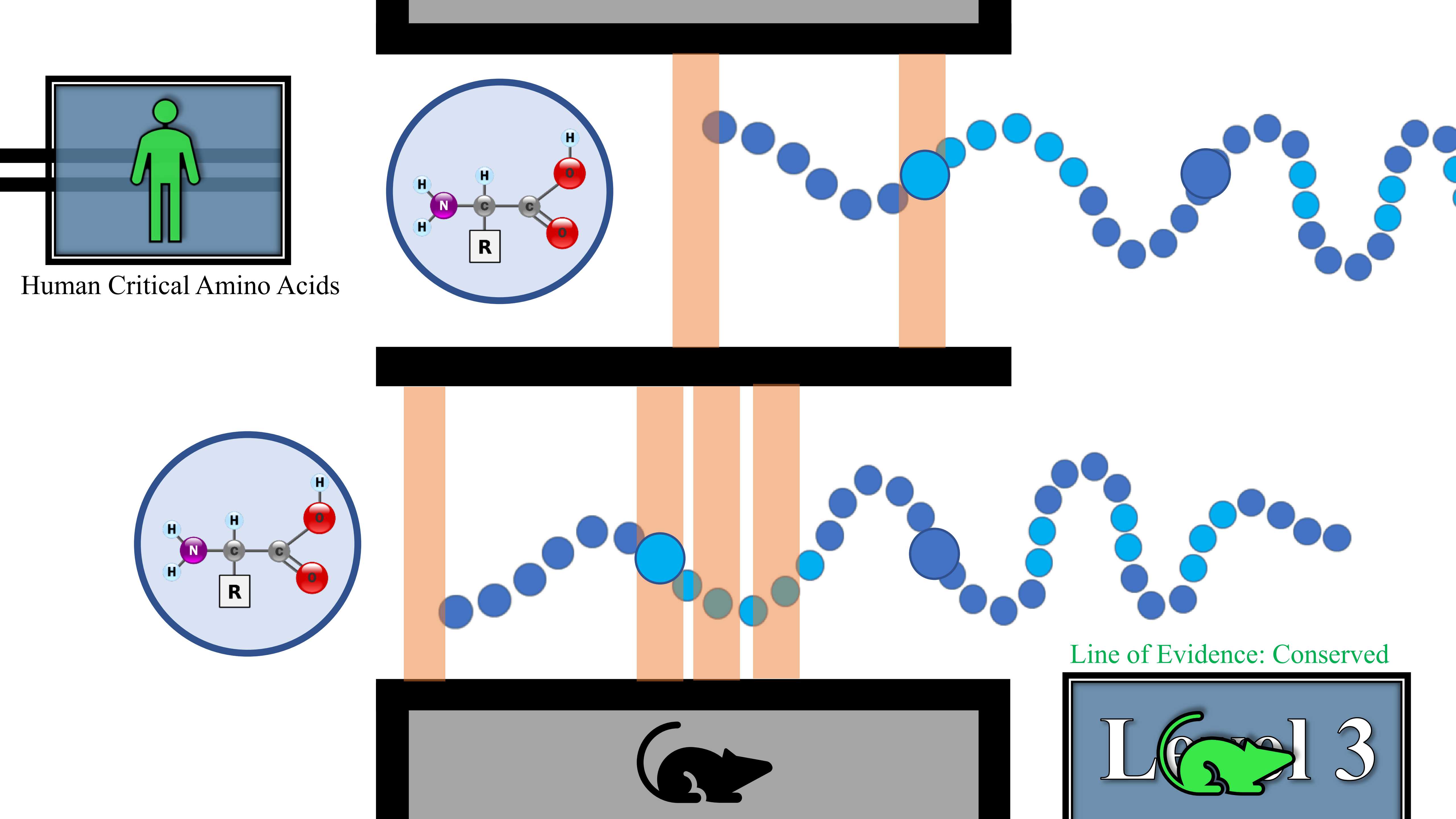
Level 2

Line of Evidence:

Domain

Not Conserved

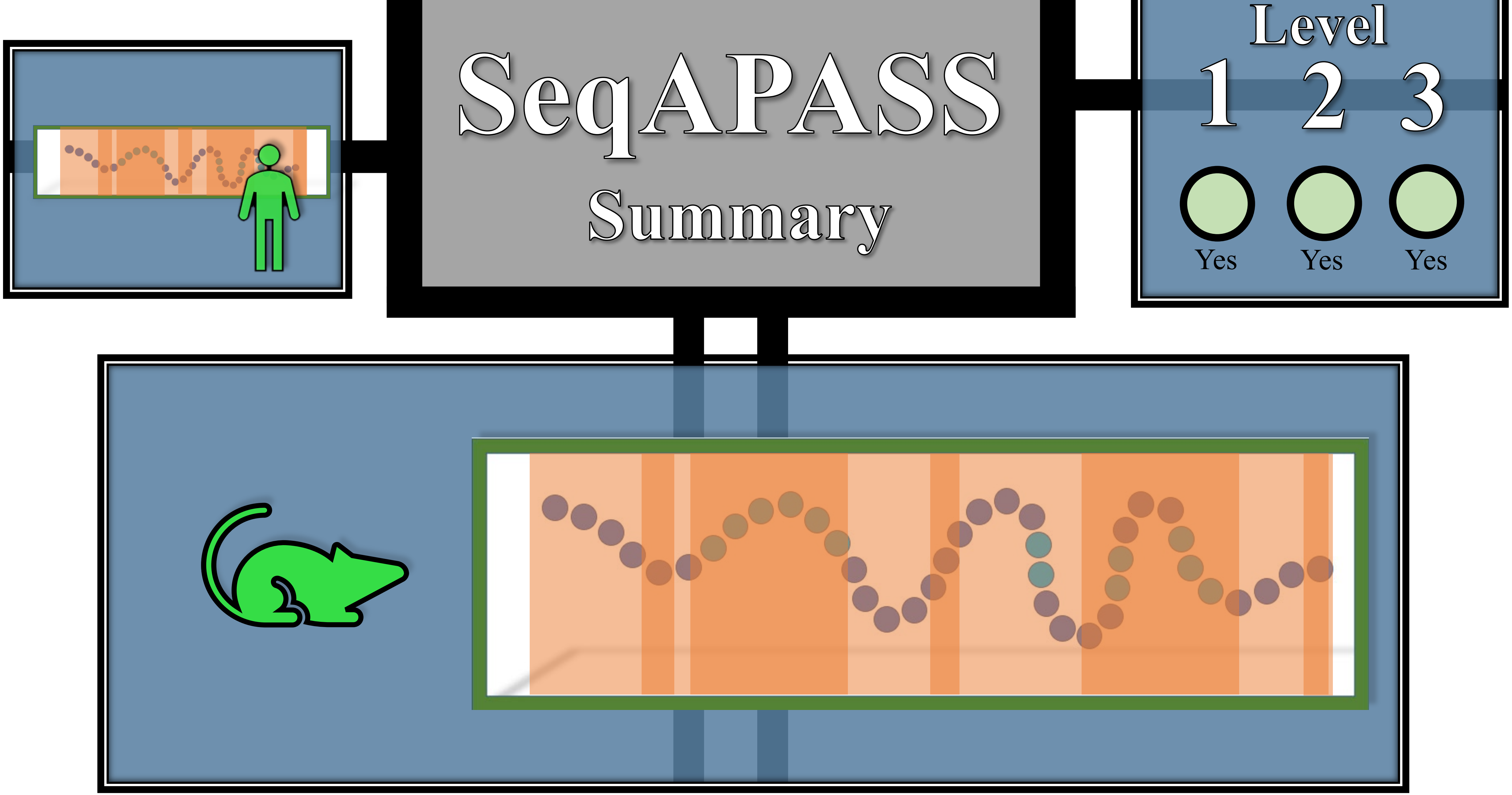




Human Critical Amino Acids

Line of Evidence: Conserved

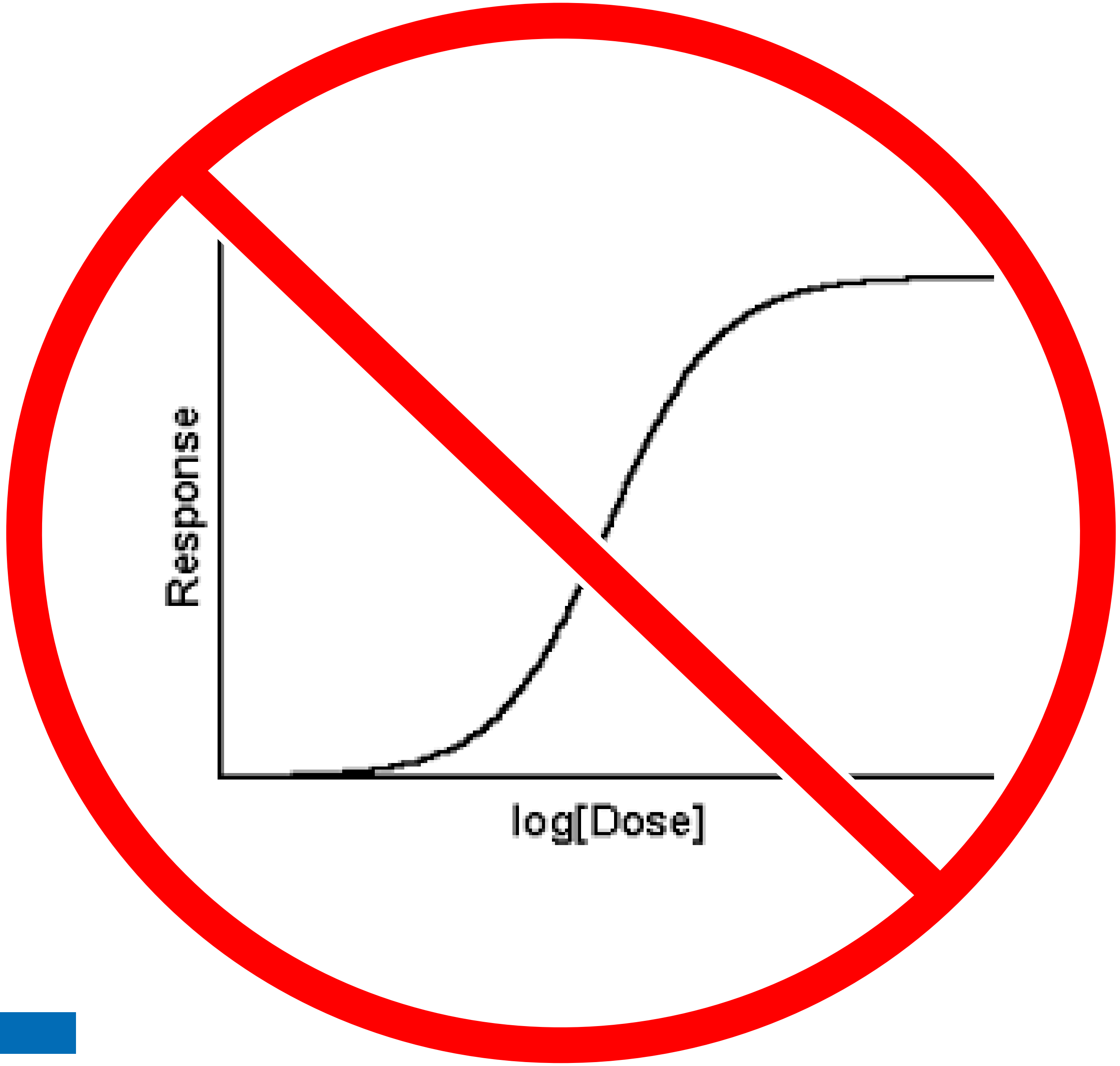
Level 3



SeqAPASS
Summary

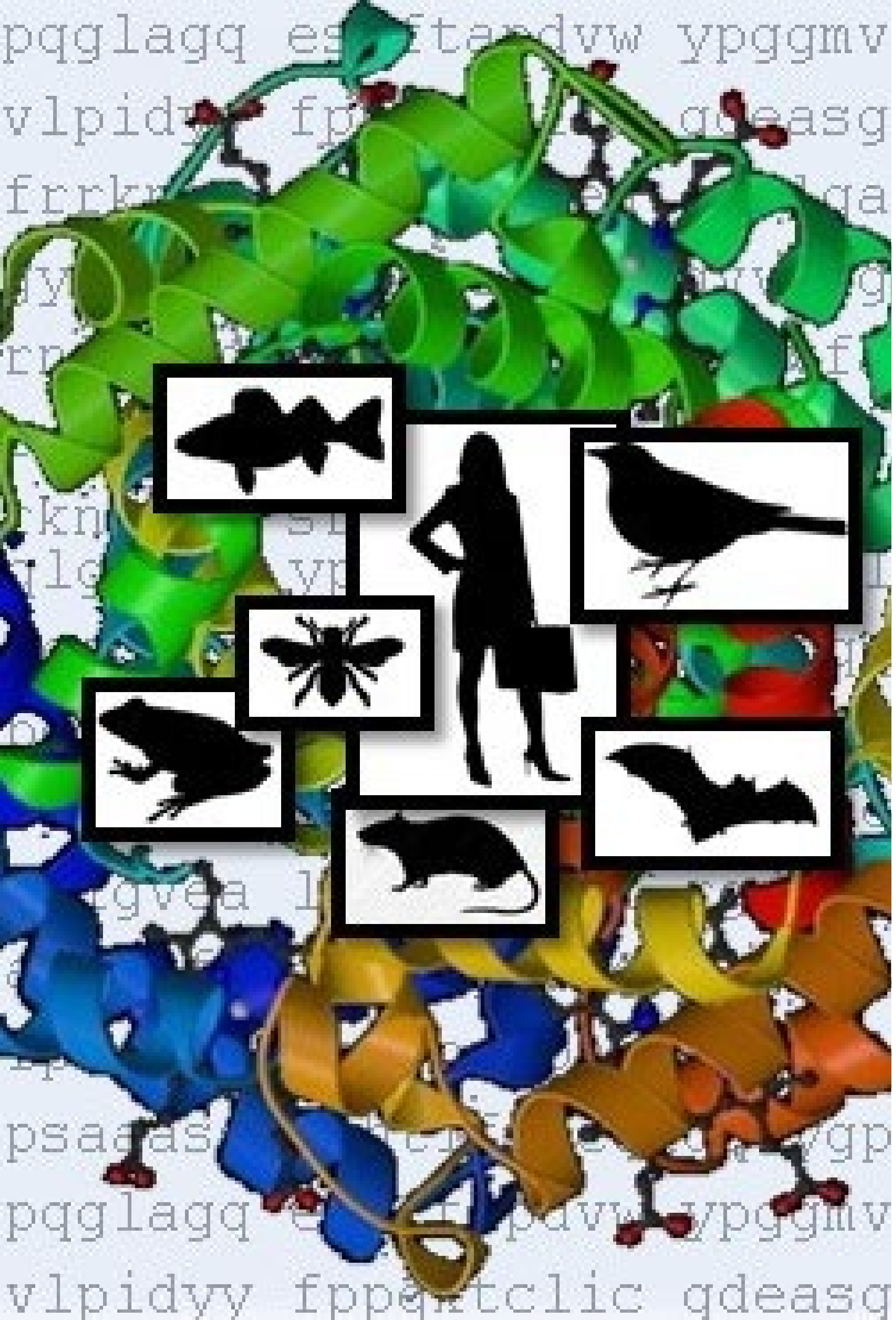
**Gather Lines of Evidence for Conservation of Protein Target:
Susceptibility Prediction: Yes or No**

SeqAPASS DOES NOT predict the degree of sensitivity/susceptibility:



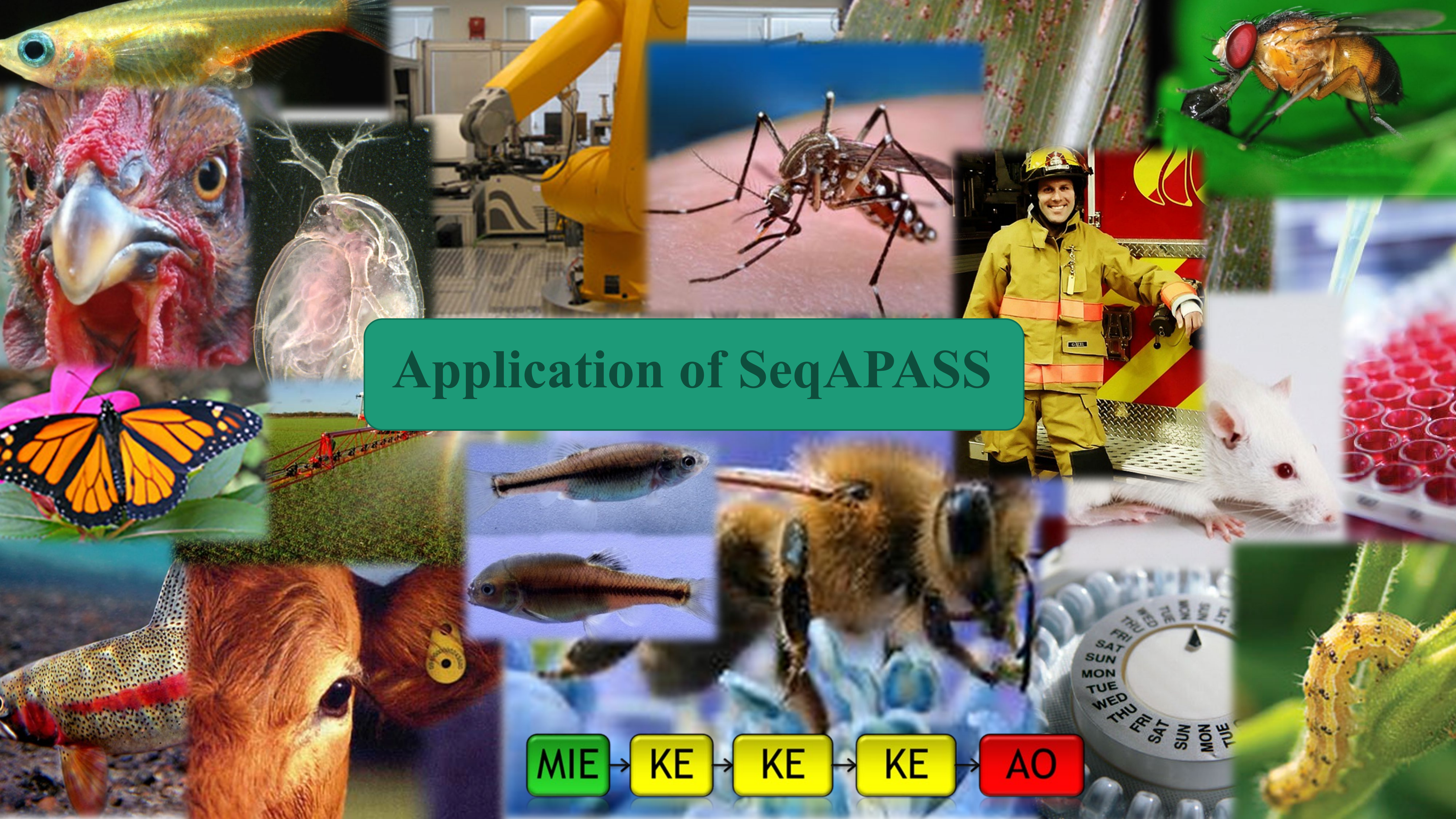
Factors that make a species sensitive

- Exposure
- Dose
- ADME
- Target receptor availability
- Life stage
- Life history
- etc.
- etc.



Strengths of SeqAPASS

- **Publicly available** to all
- Lines of evidence for conservation for **100s-1000s of species** rapidly
- Takes advantage of **well-established tools and databases**
- Streamlined, consistent, transparent, and published methods
  - **Case examples** to demonstrate applications
- **Guides users** to appropriate input
- **Evolves** as bioinformatics approaches become more user friendly
  - Smart automation or semi-automation

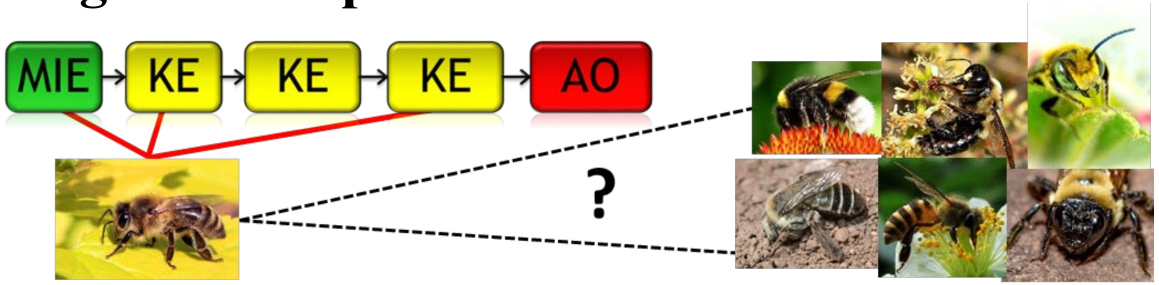


Application of SeqAPASS

MIE → KE → KE → KE → AO

Applications of Bioinformatics: Case Studies

- **Extrapolate adverse outcome pathway knowledge across species**
  - Define the taxonomic domain of applicability
    - Apis vs Non-Apis bees

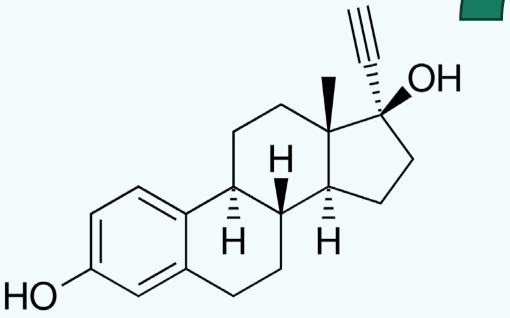


- **Extrapolate high throughput screening data**
  - Chemicals that target human estrogen receptor alpha, androgen receptor, steroidogenic enzymes, thyroid axis proteins
  - All ToxCast Assay targets
- **Predict relative intrinsic susceptibility**
  - Pesticides
  - Endangered Species Act
  - Derivation of Aquatic Life Criteria
- **Predict chemical bioaccumulation across species**
  - Chemicals of concern: PFAS
- **Generate research hypotheses** Strobilurin fungicides
- **Prioritization strategies** Pharmaceuticals

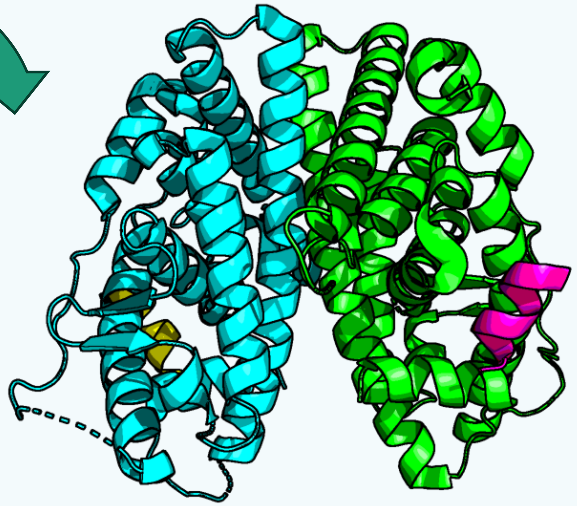
Case Study – Ethinyl Estradiol

1) **Protein**

Binds



Ethinyl estradiol



Estrogen Receptor Alpha
(ESR1; NCBI Accession: NP_000116.2)

2) **Sensitive species**

Humans are sensitive

Challenge: What other species might be *susceptible* to ethinyl estradiol?

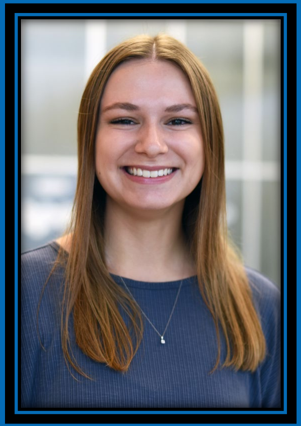
United States Environmental Protection Agency

Training Course – SeqAPASS v7.1 Level 1: Primary sequence similarity and ortholog candidate identification

Instructors:

Carlie A. LaLone, Ph.D. (US EPA)

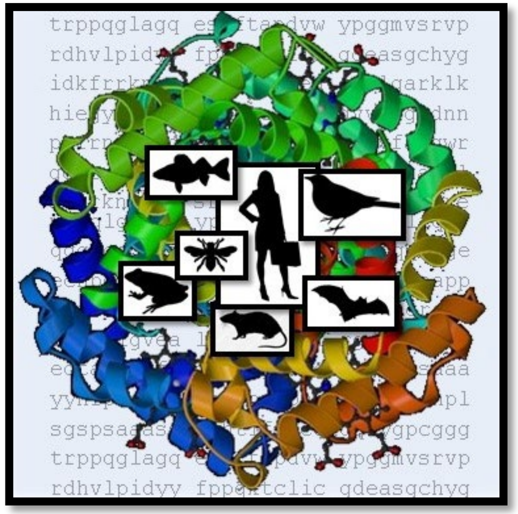
Marissa Brickley, PhD student (UMN-Duluth/ US EPA)



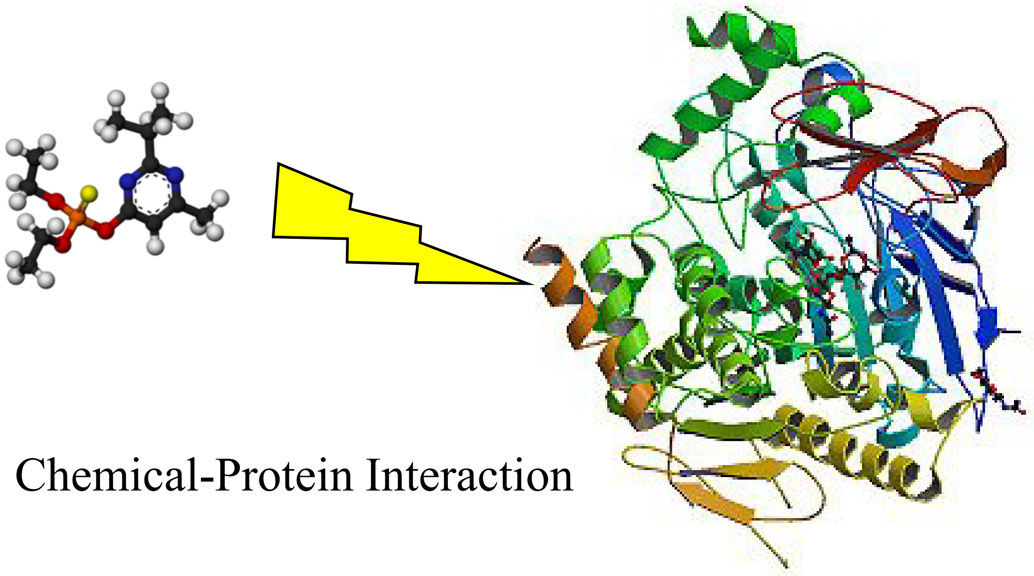
Office of Research and Development
Center for Computational Toxicology and Exposure, Great Lakes Toxicology and Ecology Division

Training Course Overview

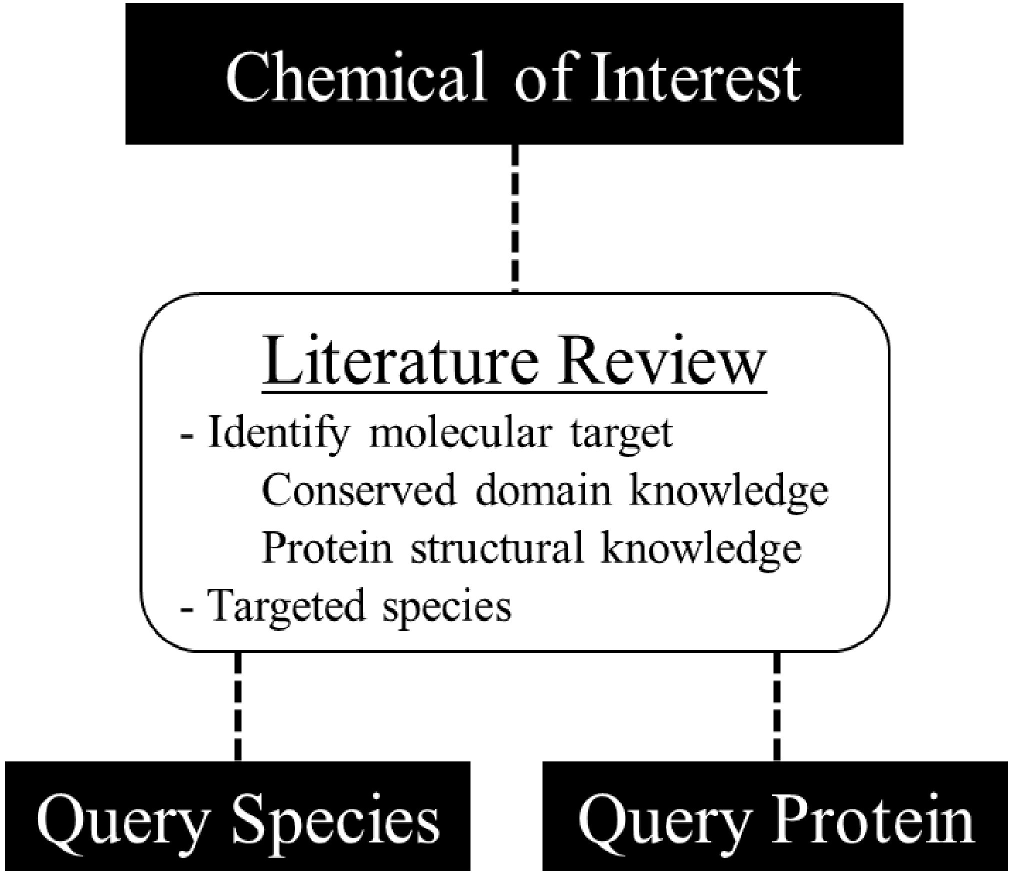
- **Introduction** to challenges in species extrapolation and how bioinformatics can aid in addressing these challenges (LaLone)
- Training with SeqAPASS
  - Level 1 and Level 1 Data visualization (LaLone)
  - Level 2 and Level 2 Data visualization (Jensen)
    - Instructor Demos – Level 1 and 2
  - Level 3 and Level 3 Data visualization (Jensen)
    - Instructor Demos – Level 3
  - Decision summary report (Jensen)
  - Level 4 (LaLone)



Query Formulation Stage



Chemical-Protein Interaction



Chemical of Interest

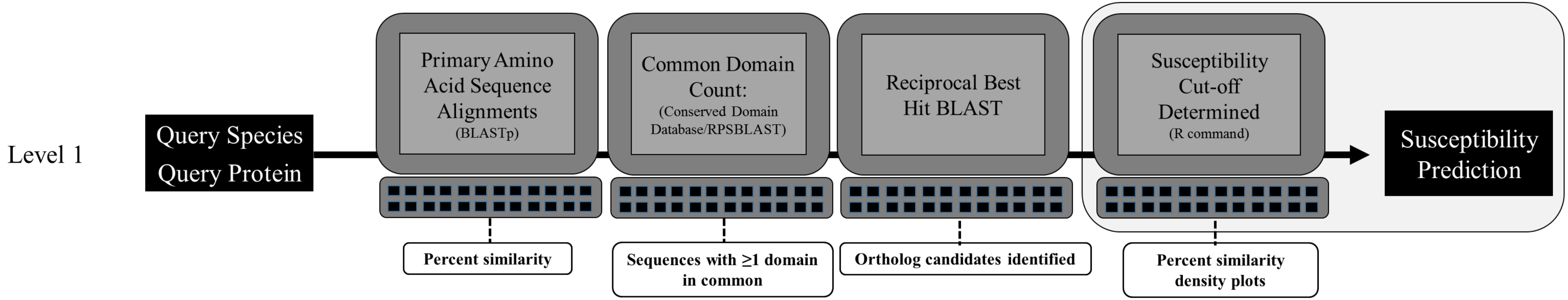
Literature Review

- Identify molecular target
  - Conserved domain knowledge
  - Protein structural knowledge
- Targeted species

Query Species

Query Protein

Level 1: Primary Amino Acid Sequence Comparisons



Level 1

Query Species
Query Protein

Primary Amino Acid Sequence Alignments
(BLASTp)

Common Domain Count:
(Conserved Domain Database/RPSBLAST)

Reciprocal Best Hit BLAST

Susceptibility Cut-off Determined
(R command)

Susceptibility Prediction

Percent similarity

Sequences with ≥1 domain in common

Ortholog candidates identified

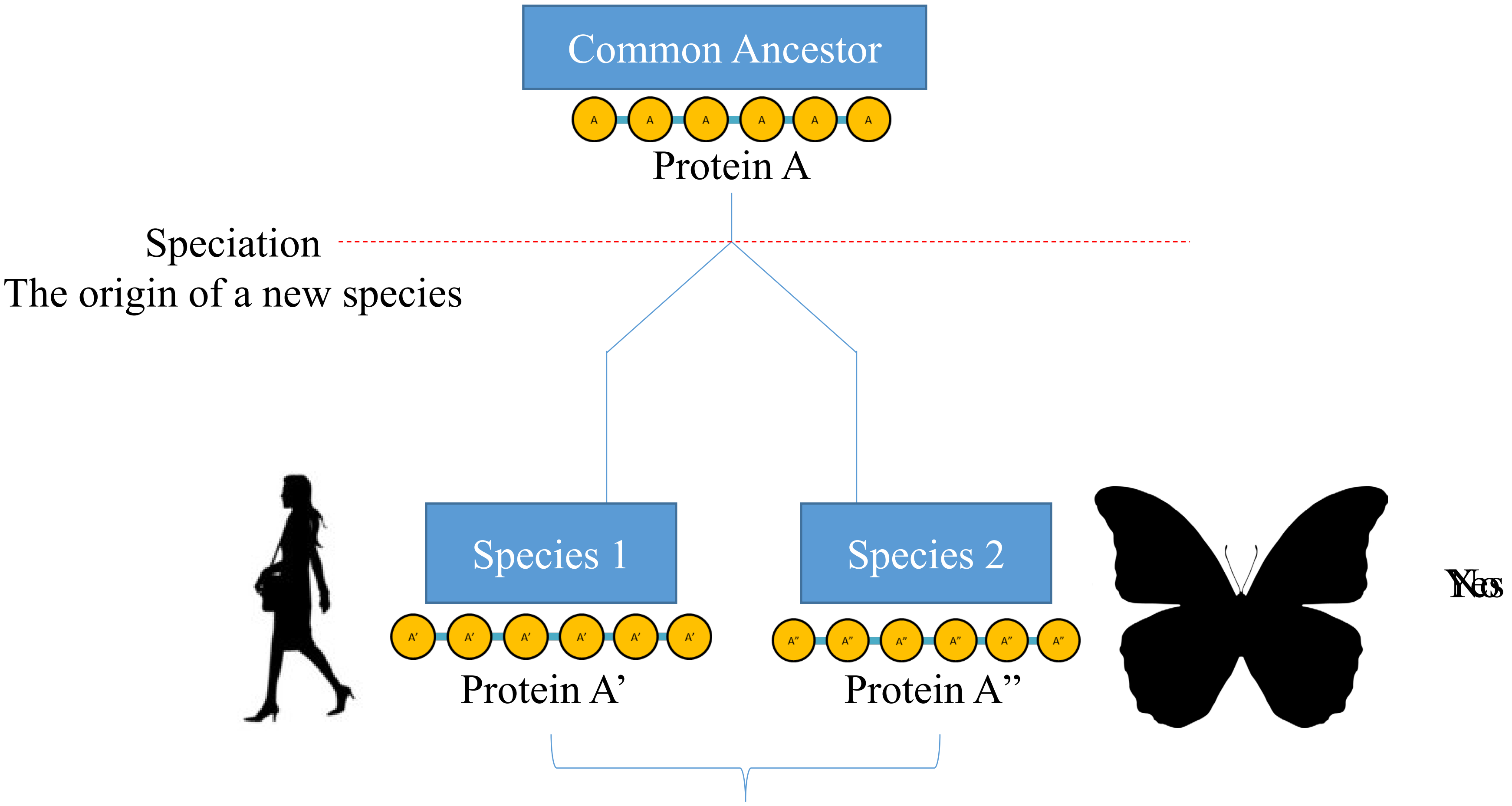
Percent similarity density plots

Align Full Primary Amino Acid Sequence

	Bit Score	Percent Similarity
Query Sequence	1241.9	100
Hit	1229.5	99.0
Hit	1223.0	98.5
Hit	1111.3	89.5
Hit	862.4	69.4

$$\text{Percent Similarity} = \frac{\text{Hit Bit Score}}{\text{Query Bit Score}} \times 100$$

Evolutionary Biology: Ortholog Candidate



Common Ancestor

Protein A

Speciation
The origin of a new species

Species 1

Species 2

Protein A’

Protein A”

Orthologs
Proteins in different species that evolved from a common ancestor

Normally, orthologs retain the same **FUNCTION** in the course of evolution

SeqAPASS Level 1 – Ortholog candidate cut-off

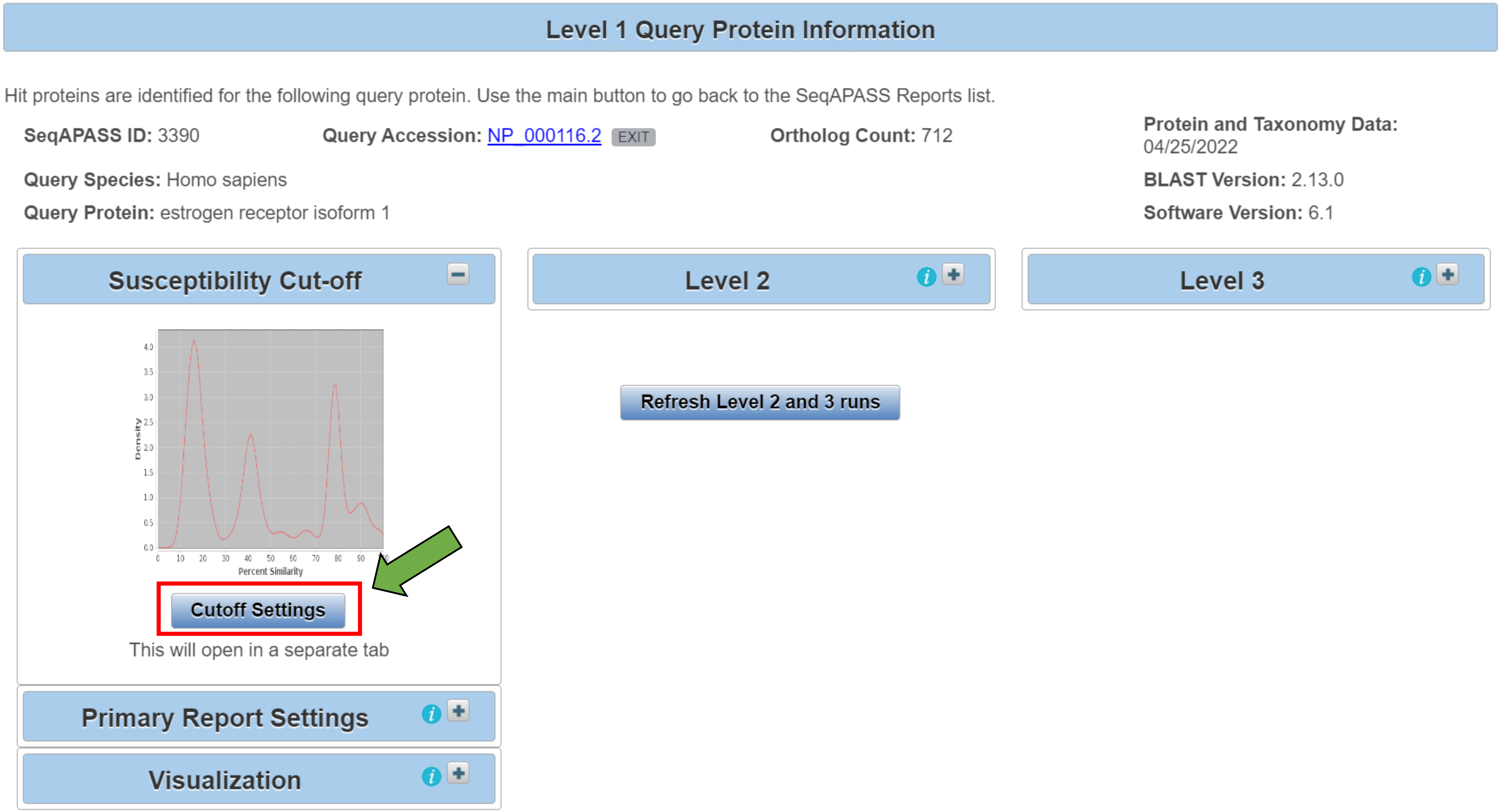
Common Name	Ortholog Candidate	Cut-off	Percent Similarity
Human	Y	33.15	100
Florida manatee	Y	33.15	98.8
Mallard	Y	33.15	82.29
Rock pigeon	Y	33.15	80.93
Green anole	Y	33.15	80.65
Pacific transparent sea squirt	Y	33.15	33.15
Yesso scallop	N	33.15	32.87
Purple sea urchin	N	33.15	26.05
Human whipworm	N	33.15	23.53
Bed bug	N	33.15	21.62

Lowest % Similarity that is still an ortholog

Example:

Susceptibility Cut-off: Set at 33.15

Above cut-off: More likely to be susceptible base on similar **FUNCTION**



Level 1 Query Protein Information

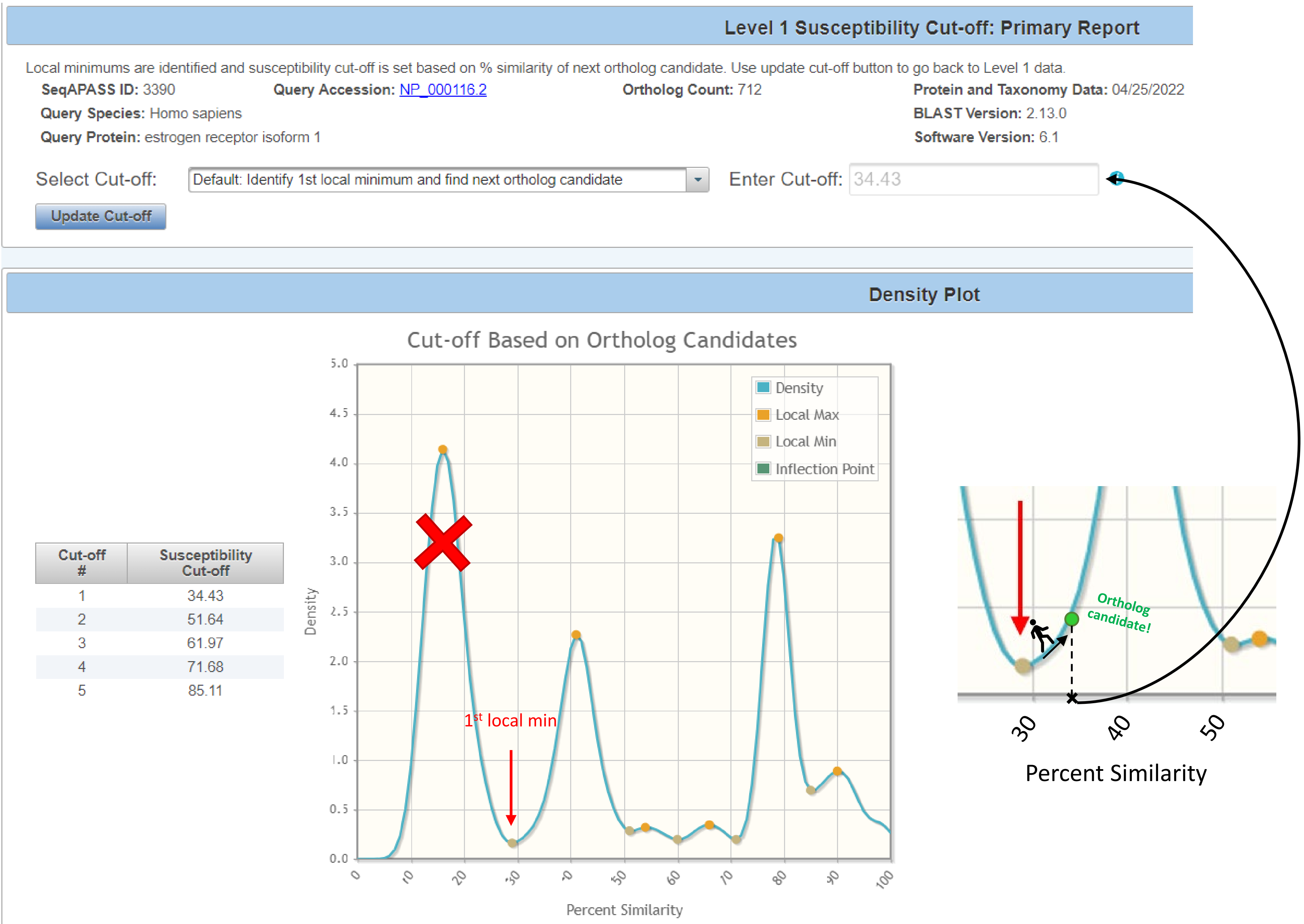
Hit proteins are identified for the following query protein. Use the main button to go back to the SeqAPASS Reports list.

SeqAPASS ID: 3390
Query Accession: NP_000116.2 EXIT
Ortholog Count: 712
Protein and Taxonomy Data: 04/25/2022

Query Species: Homo sapiens
BLAST Version: 2.13.0

Query Protein: estrogen receptor isoform 1
Software Version: 6.1

This will open in a separate tab



Level 1 Susceptibility Cut-off: Primary Report

Local minimums are identified and susceptibility cut-off is set based on % similarity of next ortholog candidate. Use update cut-off button to go back to Level 1 data.

SeqAPASS ID: 3390 **Query Accession:** NP_000116.2 **Ortholog Count:** 712 **Protein and Taxonomy Data:** 04/25/2022

Query Species: Homo sapiens **BLAST Version:** 2.13.0

Query Protein: estrogen receptor isoform 1 **Software Version:** 6.1

Select Cut-off: Default: Identify 1st local minimum and find next ortholog candidate

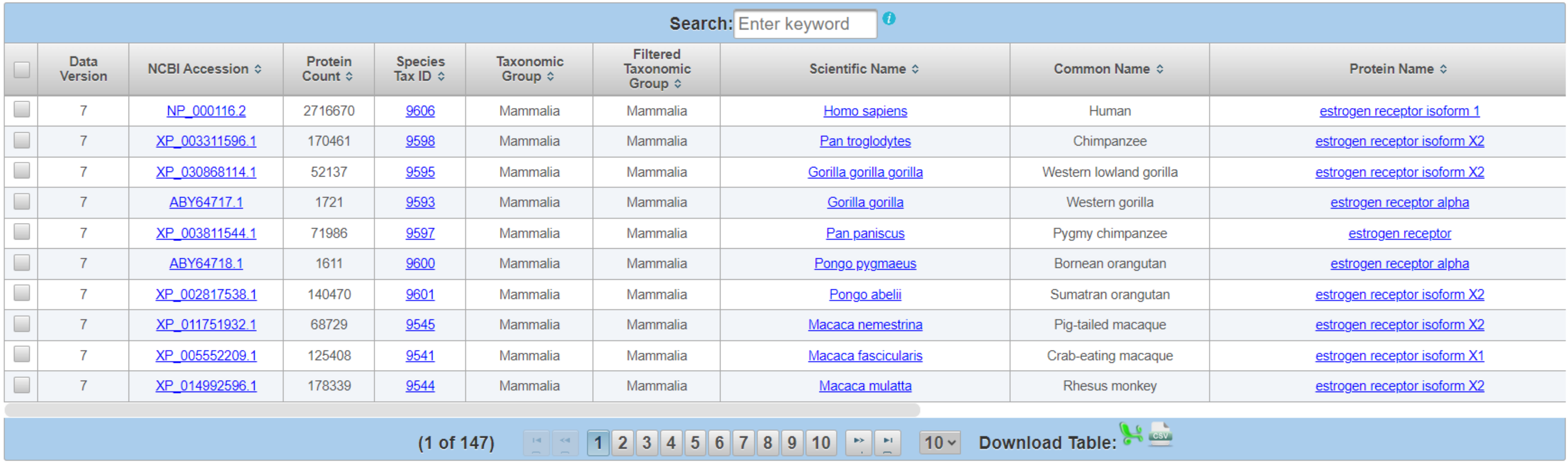
Enter Cut-off: 34.43

Density Plot

Cut-off Based on Ortholog Candidates

Cut-off #	Susceptibility Cut-off
1	34.43
2	51.64
3	61.97
4	71.68
5	85.11

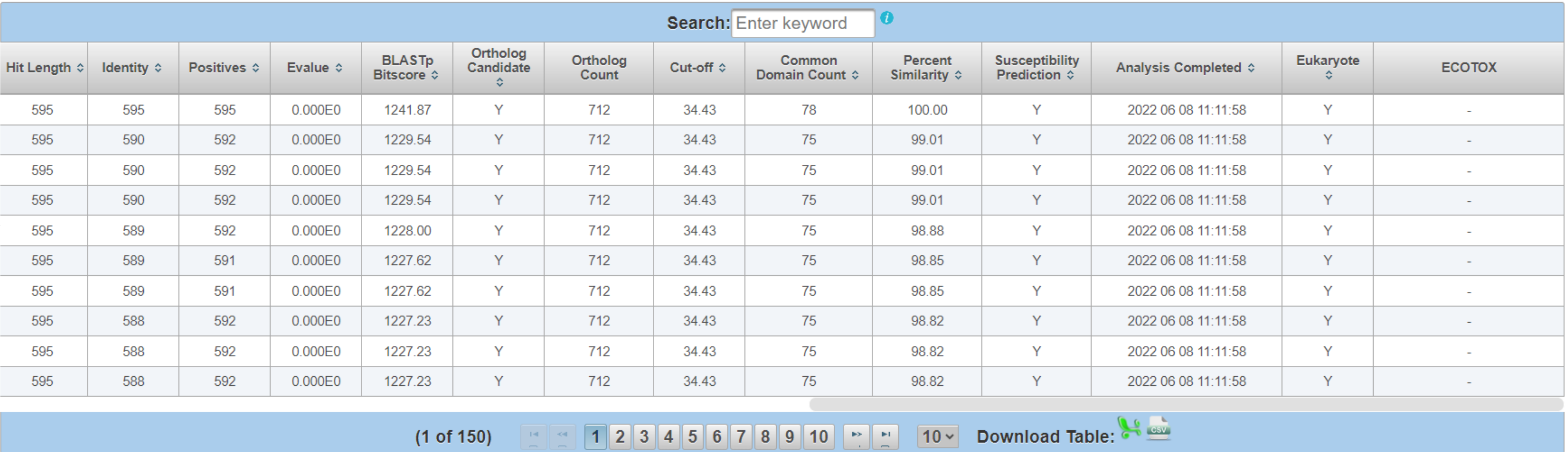
Level 1 SeqAPASS Data



Search: Enter keyword

☐	Data Version	NCBI Accession	Protein Count	Species Tax ID	Taxonomic Group	Filtered Taxonomic Group	Scientific Name	Common Name	Protein Name
☐	7	NP_000116.2	2716670	9606	Mammalia	Mammalia	Homo sapiens	Human	estrogen receptor isoform 1
☐	7	XP_003311596.1	170461	9598	Mammalia	Mammalia	Pan troglodytes	Chimpanzee	estrogen receptor isoform X2
☐	7	XP_030868114.1	52137	9595	Mammalia	Mammalia	Gorilla gorilla gorilla	Western lowland gorilla	estrogen receptor isoform X2
☐	7	ABY64717.1	1721	9593	Mammalia	Mammalia	Gorilla gorilla	Western gorilla	estrogen receptor alpha
☐	7	XP_003811544.1	71986	9597	Mammalia	Mammalia	Pan paniscus	Pygmy chimpanzee	estrogen receptor
☐	7	ABY64718.1	1611	9600	Mammalia	Mammalia	Pongo pygmaeus	Bornean orangutan	estrogen receptor alpha
☐	7	XP_002817538.1	140470	9601	Mammalia	Mammalia	Pongo abelii	Sumatran orangutan	estrogen receptor isoform X2
☐	7	XP_011751932.1	68729	9545	Mammalia	Mammalia	Macaca nemestrina	Pig-tailed macaque	estrogen receptor isoform X2
☐	7	XP_005552209.1	125408	9541	Mammalia	Mammalia	Macaca fascicularis	Crab-eating macaque	estrogen receptor isoform X1
☐	7	XP_014992596.1	178339	9544	Mammalia	Mammalia	Macaca mulatta	Rhesus monkey	estrogen receptor isoform X2

(1 of 147) 1 2 3 4 5 6 7 8 9 10 10 **Download Table:**

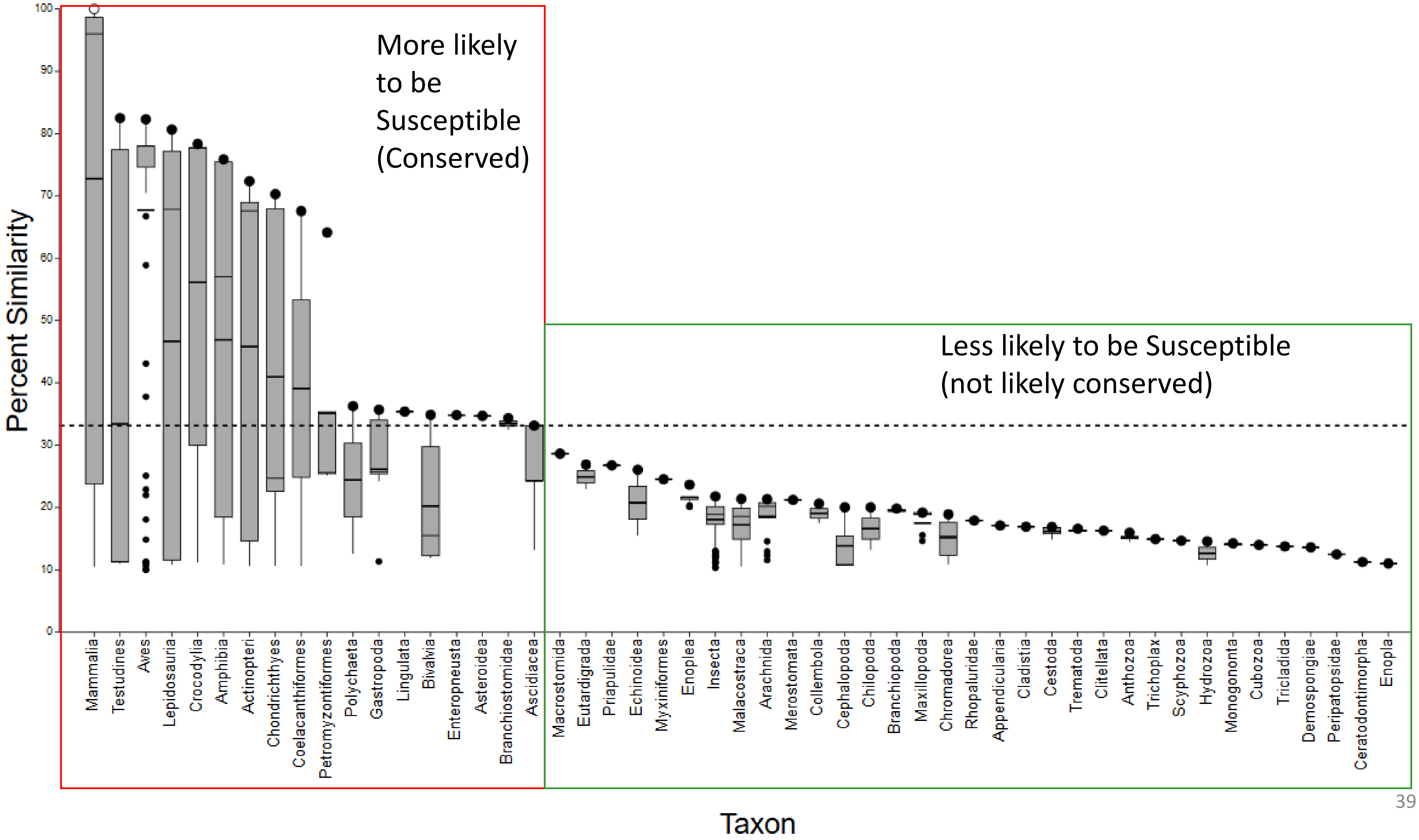


Search: Enter keyword

Hit Length	Identity	Positives	Evalue	BLASTp Bitscore	Ortholog Candidate	Ortholog Count	Cut-off	Common Domain Count	Percent Similarity	Susceptibility Prediction	Analysis Completed	Eukaryote	ECOTOX
595	595	595	0.000E0	1241.87	Y	712	34.43	78	100.00	Y	2022 06 08 11:11:58	Y	-
595	590	592	0.000E0	1229.54	Y	712	34.43	75	99.01	Y	2022 06 08 11:11:58	Y	-
595	590	592	0.000E0	1229.54	Y	712	34.43	75	99.01	Y	2022 06 08 11:11:58	Y	-
595	590	592	0.000E0	1229.54	Y	712	34.43	75	99.01	Y	2022 06 08 11:11:58	Y	-
595	589	592	0.000E0	1228.00	Y	712	34.43	75	98.88	Y	2022 06 08 11:11:58	Y	-
595	589	591	0.000E0	1227.62	Y	712	34.43	75	98.85	Y	2022 06 08 11:11:58	Y	-
595	589	591	0.000E0	1227.62	Y	712	34.43	75	98.85	Y	2022 06 08 11:11:58	Y	-
595	588	592	0.000E0	1227.23	Y	712	34.43	75	98.82	Y	2022 06 08 11:11:58	Y	-
595	588	592	0.000E0	1227.23	Y	712	34.43	75	98.82	Y	2022 06 08 11:11:58	Y	-
595	588	592	0.000E0	1227.23	Y	712	34.43	75	98.82	Y	2022 06 08 11:11:58	Y	-

(1 of 150) 1 2 3 4 5 6 7 8 9 10 10 **Download Table:**

Box-Plot of SeqAPASS Data



More likely
to be
Susceptible
(Conserved)

Less likely to be Susceptible
(not likely conserved)

United States
Environmental Protection
Agency

Training Course - SeqAPASS v7.1 Level 2: Conservation of Functional Domains

Instructors:

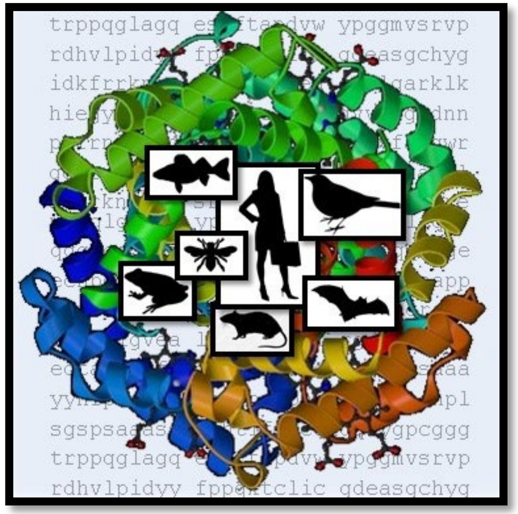
Carlie A. LaLone, Ph.D. (US EPA)

Marissa Brickley, PhD student (UMN-Duluth/ US EPA)

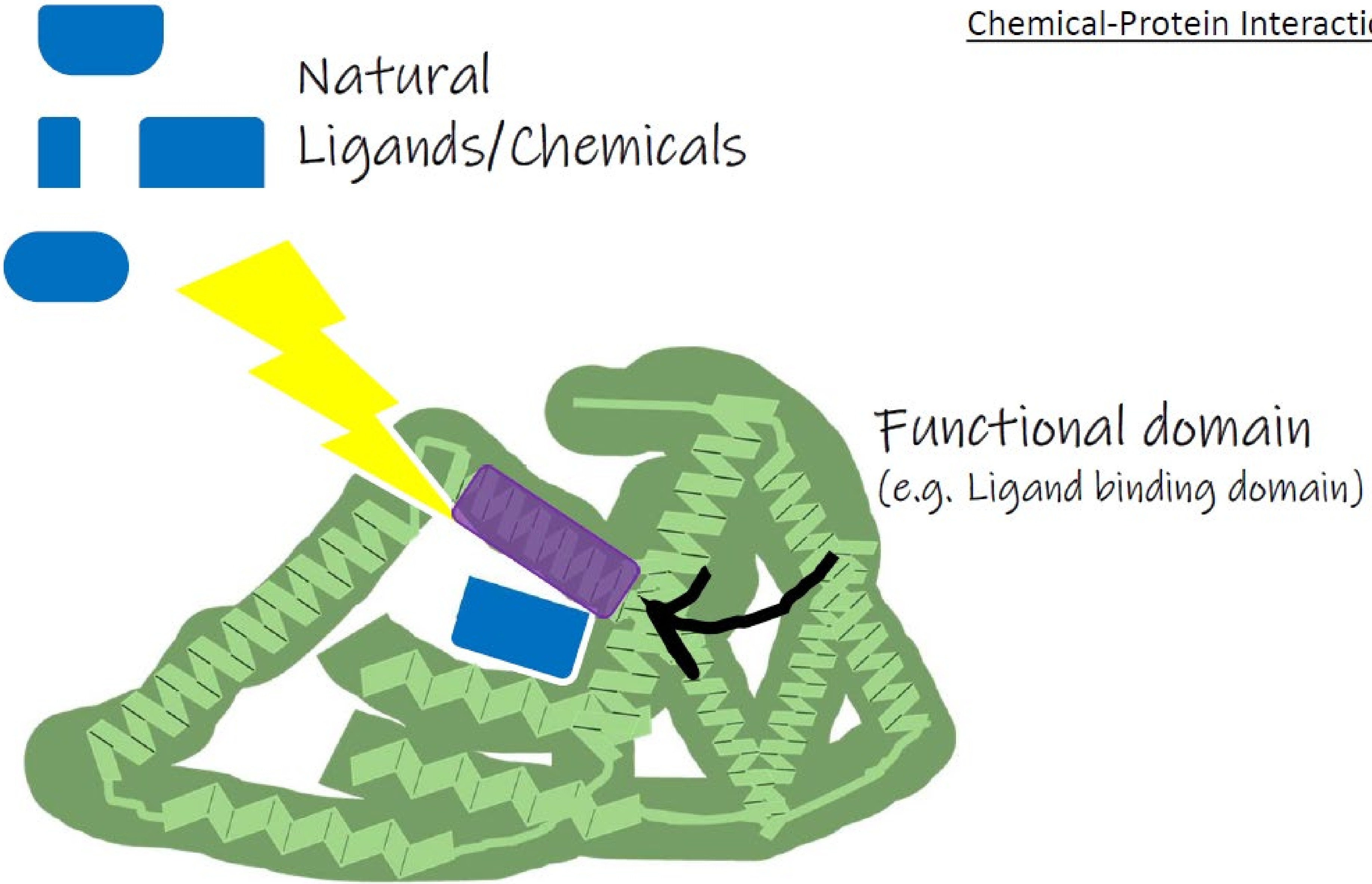
Office of Research and Development
Center for Computational Toxicology and Exposure, Great Lakes Toxicology and Ecology Division

Training Course Overview

- **Introduction** to challenges in species extrapolation and how bioinformatics can aid in addressing these challenges (LaLone)
- Training with SeqAPASS
    - Level 1 and Level 1 Data visualization (LaLone)
    - Level 2 and Level 2 Data visualization (Jensen)
        - Instructor Demos – Level 1 and 2
    - Level 3 and Level 3 Data visualization (Jensen)
        - Instructor Demos – Level 3
    - Decision summary report (Jensen)
    - Level 4 (LaLone)

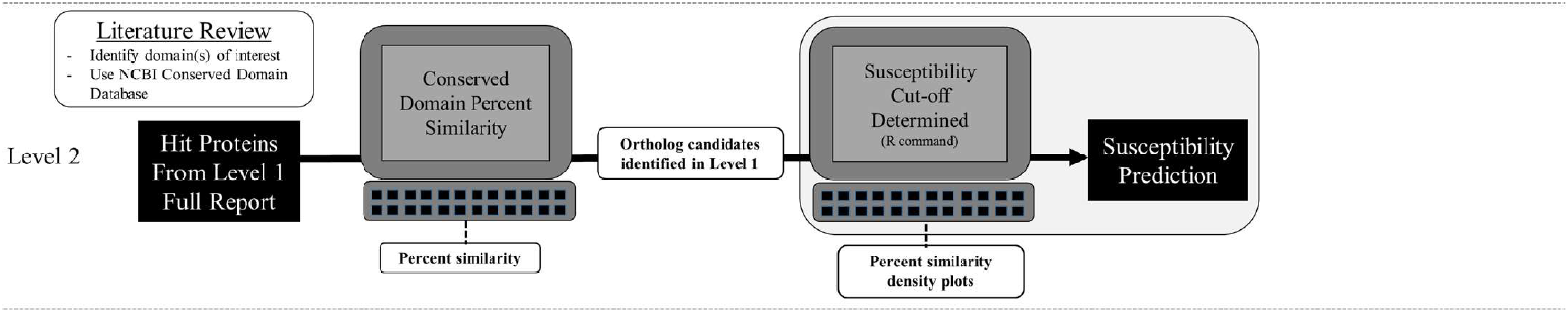


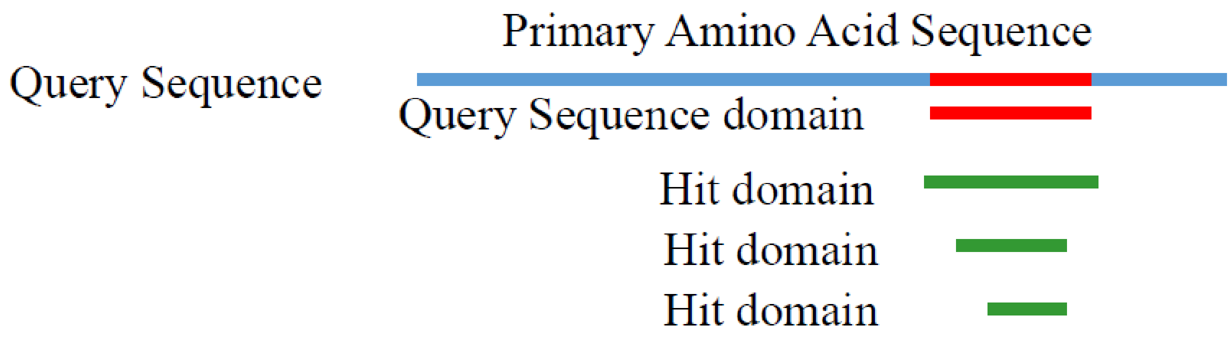
Chemical-Protein Interaction:



Similarity across species at the molecular level

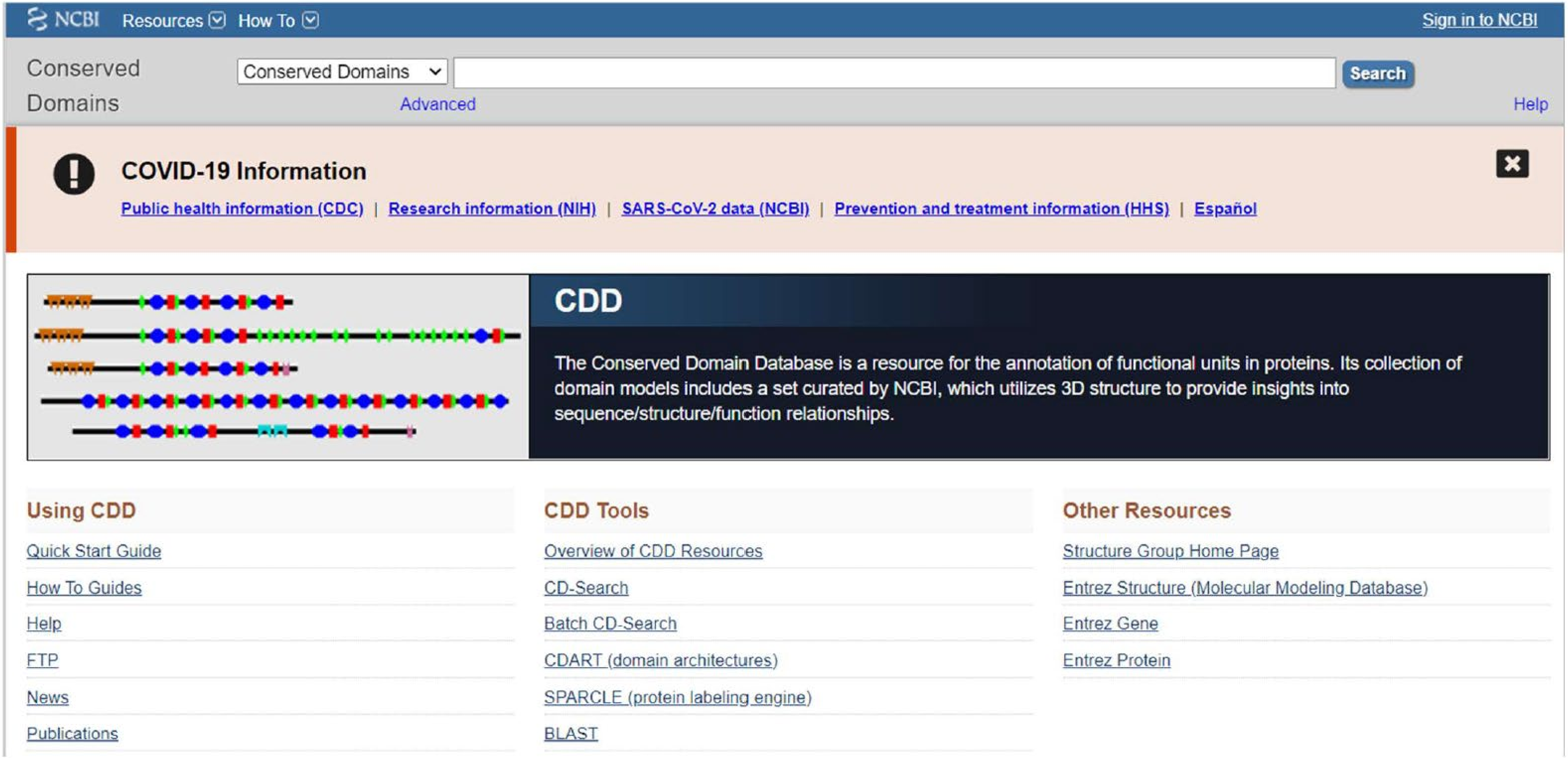
Level 2: Functional domain comparisons





Bit Score	Percent Similarity
482.6	100
471.9	97.8
303.5	62.9
100.1	20.7

$$\text{Percent Similarity} = \frac{\text{Hit Bit Score}}{\text{Query Bit Score}} \text{ X } 100$$



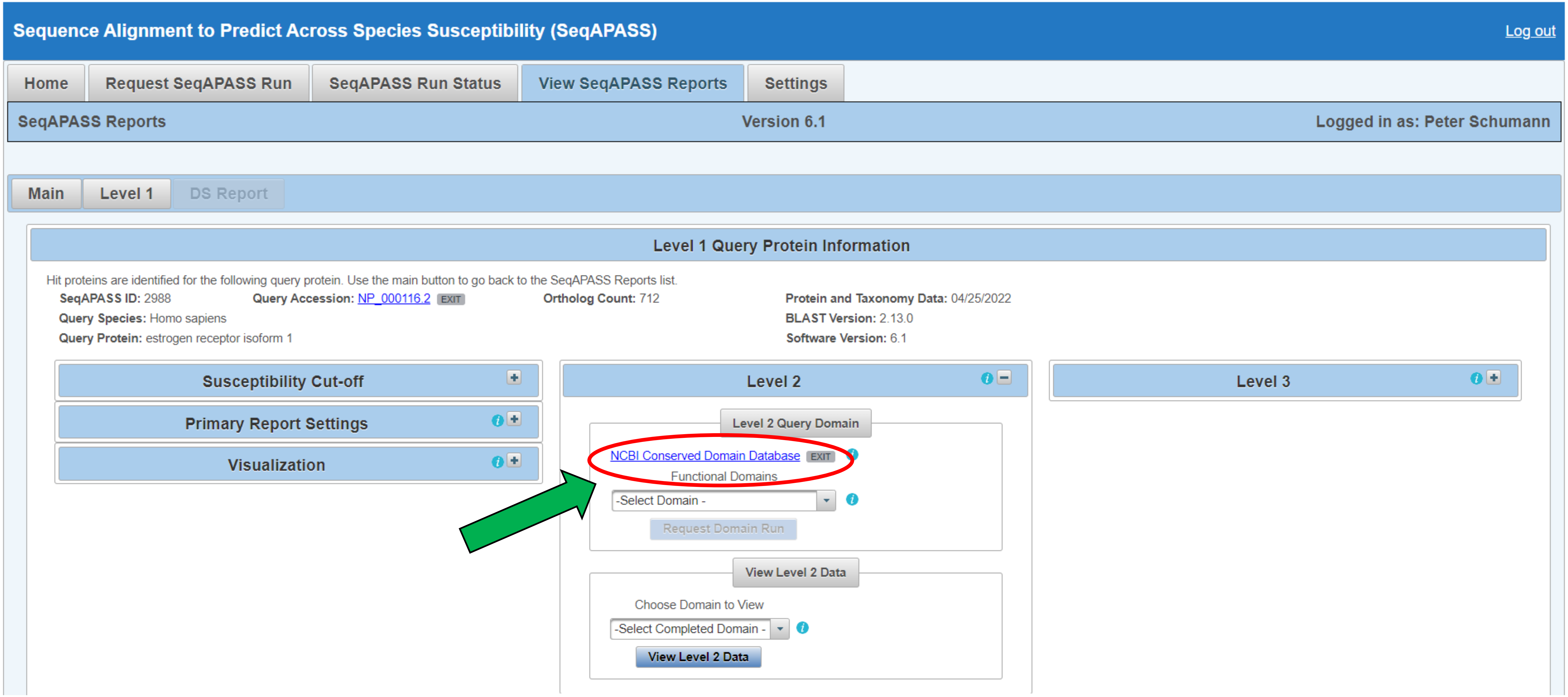
COVID-19 Information

Public health information (CDC) | Research information (NIH) | SARS-CoV-2 data (NCBI) | Prevention and treatment information (HHS) | Español

CDD

The Conserved Domain Database is a resource for the annotation of functional units in proteins. Its collection of domain models includes a set curated by NCBI, which utilizes 3D structure to provide insights into sequence/structure/function relationships.

Using CDD	CDD Tools	Other Resources
Quick Start Guide	Overview of CDD Resources	Structure Group Home Page
How To Guides	CD-Search	Entrez Structure (Molecular Modeling Database)
Help	Batch CD-Search	Entrez Gene
FTP	CDART (domain architectures)	Entrez Protein
News	SPARCLE (protein labeling engine)	
Publications	BLAST	



Hit proteins are identified for the following query protein. Use the main button to go back to the SeqAPASS Reports list.

SeqAPASS ID: 2988
Query Accession: NP_000116.2 EXIT
Ortholog Count: 712
Protein and Taxonomy Data: 04/25/2022
Query Species: Homo sapiens
BLAST Version: 2.13.0
Query Protein: estrogen receptor isoform 1
Software Version: 6.1

Conserved domains on [gi|62821794|ref|NP_000116|]

estrogen receptor isoform 1 [Homo sapiens]

Protein Classification

NR_DBD_ER and NR_LBD_ER domain-containing protein (domain architecture ID 10490022)
protein containing domains Oest_recep, NR_DBD_ER, NR_LBD_ER, and ESR1_C

Graphical summary

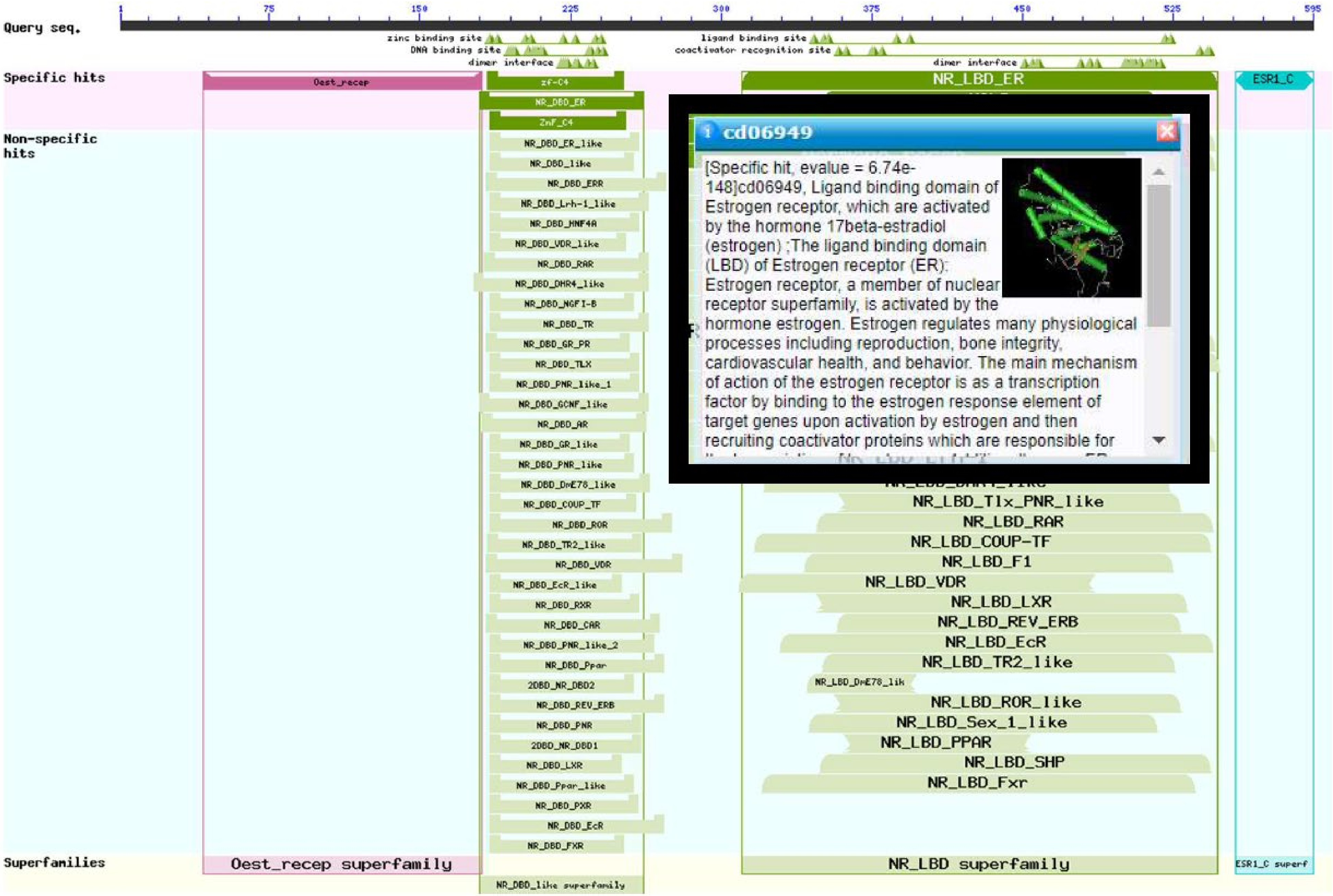
☐ Zoom to residue level show extra options »

List of domain hits

Name	Accession	Description	Interval	E-value
[+] NR_LBD_ER	cd06949	Ligand binding domain of Estrogen receptor, which are activated by the hormone ...	310-547	6.74e-148
[+] Oest_recep	pfam02159	Oestrogen receptor;	42-181	2.15e-72
[+] NR_DBD_ER	cd07171	DNA-binding domain of estrogen receptors (ER) is composed of two C4-type zinc fingers; ...	180-261	6.64e-63
[+] ESR1_C	pfam12743	Oestrogen-type nuclear receptor final C-terminal; This is the very C-terminal region of a ...	556-595	3.28e-19

References:

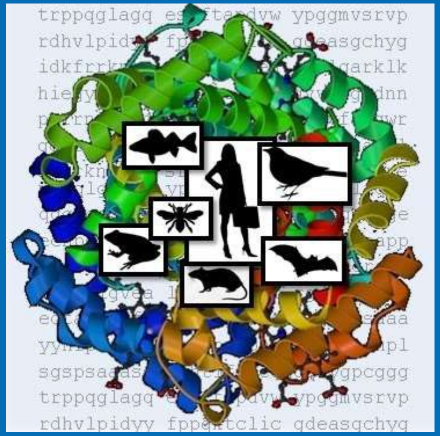
Marchler-Bauer A et al. (2017), *"CDD/SPARCLE: functional classification of proteins via subfamily domain architectures."*, **Nucleic Acids Res.45**(D)200-3.



[Specific hit, evalue = 6.74e-148]cd06949, Ligand binding domain of Estrogen receptor, which are activated by the hormone 17beta-estradiol (estrogen) ;The ligand binding domain (LBD) of Estrogen receptor (ER); Estrogen receptor, a member of nuclear receptor superfamily, is activated by the hormone estrogen. Estrogen regulates many physiological processes including reproduction, bone integrity, cardiovascular health, and behavior. The main mechanism of action of the estrogen receptor is as a transcription factor by binding to the estrogen response element of target genes upon activation by estrogen and then recruiting coactivator proteins which are responsible for

SeqAPASS Live Demo

Level 1 & 2 setup, output, and visualization



United States
Environmental Protection
Agency

Training Course – SeqAPASS v7.1 Level 3: Critical individual amino acid comparison

Instructors:

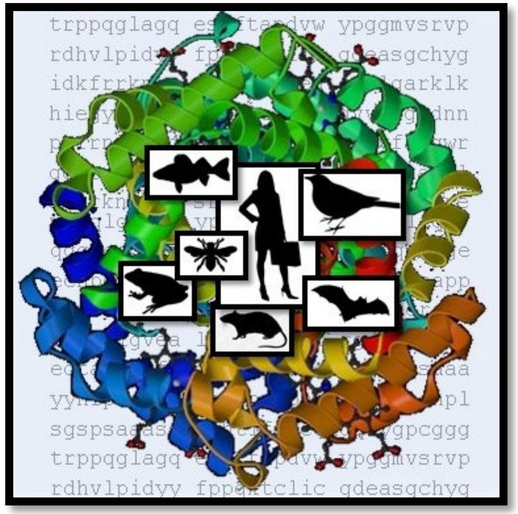
Carlie A. LaLone, Ph.D. (US EPA)

Marissa Brickley, PhD student (UMN-Duluth/ US EPA)

Office of Research and Development
Center for Computational Toxicology and Exposure, Great Lakes Toxicology and Ecology Division

Training Course Overview

- **Introduction** to challenges in species extrapolation and how bioinformatics can aid in addressing these challenges (LaLone)
- Training with SeqAPASS
  - Level 1 and Level 1 Data visualization (LaLone)
  - Level 2 and Level 2 Data visualization (Jensen)
    - Instructor Demos – Level 1 and 2
  - Level 3 and Level 3 Data visualization (Jensen)
    - Instructor Demos – Level 3
  - Decision summary report (Jensen)
  - Level 4 (LaLone)

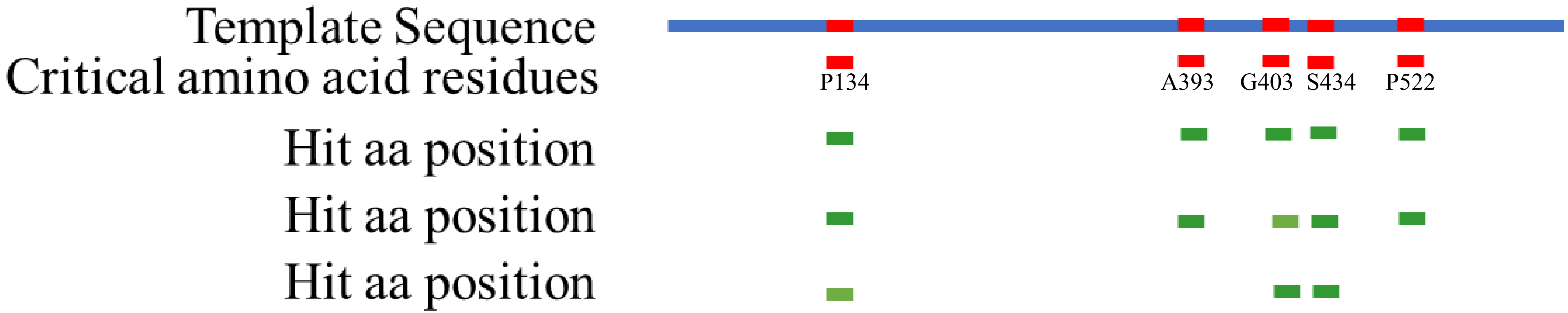


Level 3: Individual Amino Acid Residue Comparison

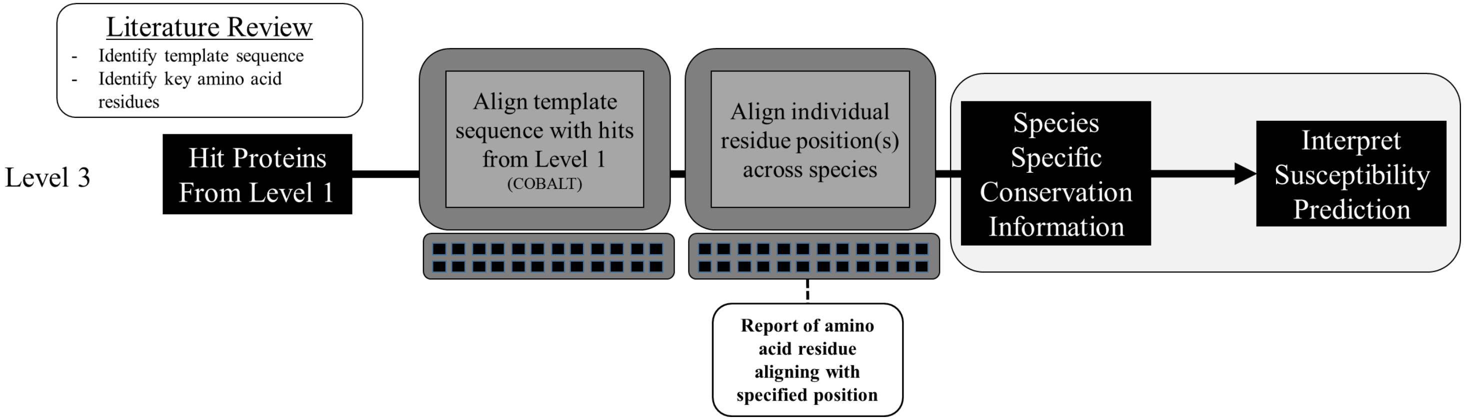
Necessary information to begin Level 3 query

1. Template Species/Sequence

2. Identified Critical Amino Acids and positions of amino acids in template

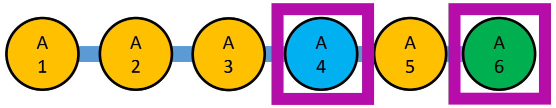


Level 3: Individual Amino Acid Residue Comparison



Crystal structure of protein – Ligand bound

Amino acid residues that interact with the chemical

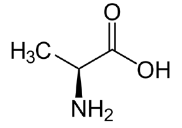


Where can I find this information?

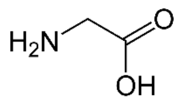
- Literature Review
  - Types of studies:
    - Site-directed mutagenesis
    - Field resistance (pesticides)
    - Studies of x-ray crystallography
    - Homology modeling

Level 3: Continued

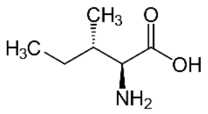
Aliphatic



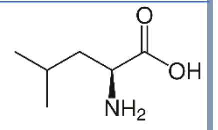
Alanine (A)
MW: 89.094



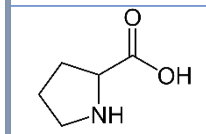
Glycine (G)
MW: 75.067



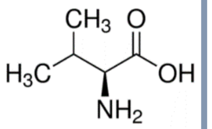
Isoleucine (I)
MW: 131.175



Leucine (L)
MW: 131.175

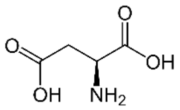


Proline (P)
MW: 115.132

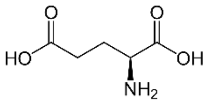


Valine (V)
MW: 117.148

Acidic

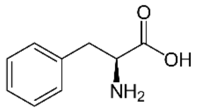


Aspartic Acid (D)
MW: 133.104

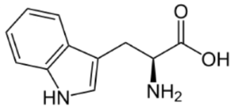


Glutamic Acid (E)
MW: 147.131

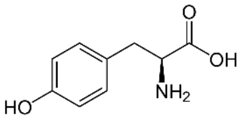
Aromatic



Phenylalanine (F)
MW: 165.192

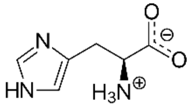


Tryptophan (W)
MW: 204.228

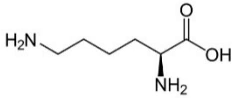


Tyrosine (Y)
MW: 181.191

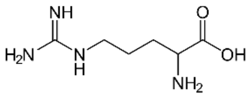
Basic



Histidine (H)
MW: 155.156

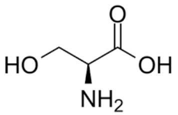


Lysine (K)
MW: 146.189

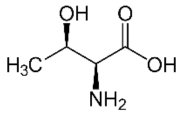


Arginine (R)
MW: 174.203

Hydroxylic

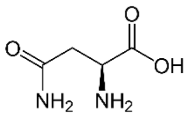


Serine (S)
MW: 105.093

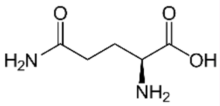


Threonine (T)
MW: 119.119

Amidic

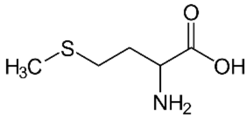


Asparagine (N)
MW: 132.119

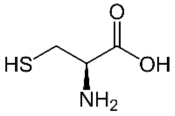


Glutamine (Q)
MW: 146.146

Sulfur-Containing



Methionine (M)
MW: 149.208



Cysteine (C)
MW: 121.154

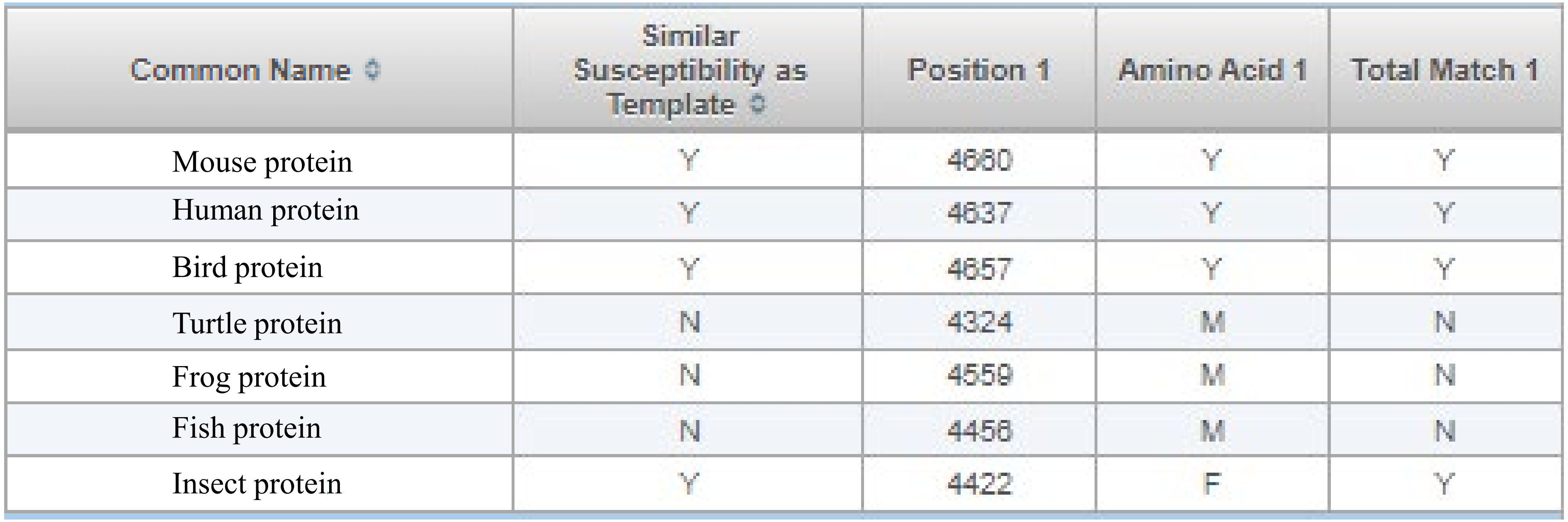
SeqAPASS can **AUTOMATICALLY** predict whether an amino acid difference is likely to change protein-chemical interaction.

- Automated Prediction
  - Each key amino acid is grouped into a "side chain class" and given a "size"
- Evaluated based on Rules:
  - Same side chain class as query (Y/N)?
  - Size 30g/mol or less from query (Y/N)?
- If 2 "N" responses for ≥1 key amino acids then SeqAPASS predicts susceptibility of that species will differ from the query species.

Level 3: Continued

Example of how Level 3 Susceptibility Predictions work (hypothetical case study):

Primary report:

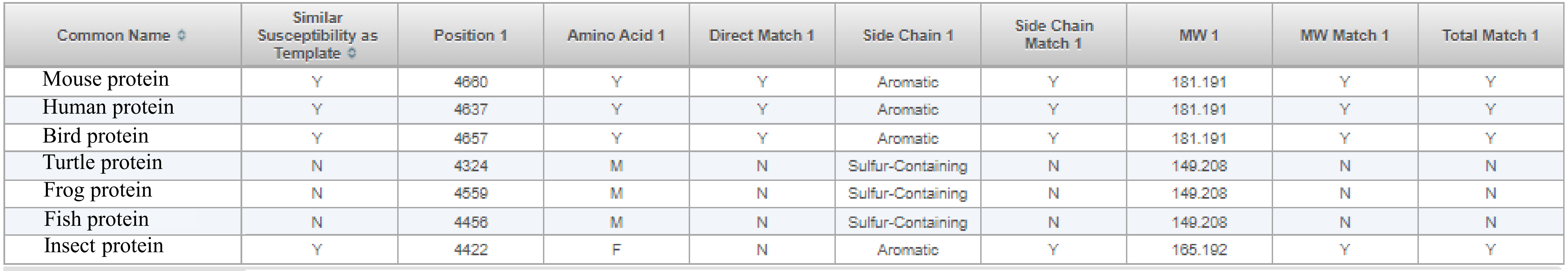


Common Name	Similar Susceptibility as Template	Position 1	Amino Acid 1	Total Match 1
Mouse protein	Y	4660	Y	Y
Human protein	Y	4637	Y	Y
Bird protein	Y	4657	Y	Y
Turtle protein	N	4324	M	N
Frog protein	N	4559	M	N
Fish protein	N	4456	M	N
Insect protein	Y	4422	F	Y

Level 3: Continued

Example of how Level 3 Susceptibility Predictions work (hypothetical case study):

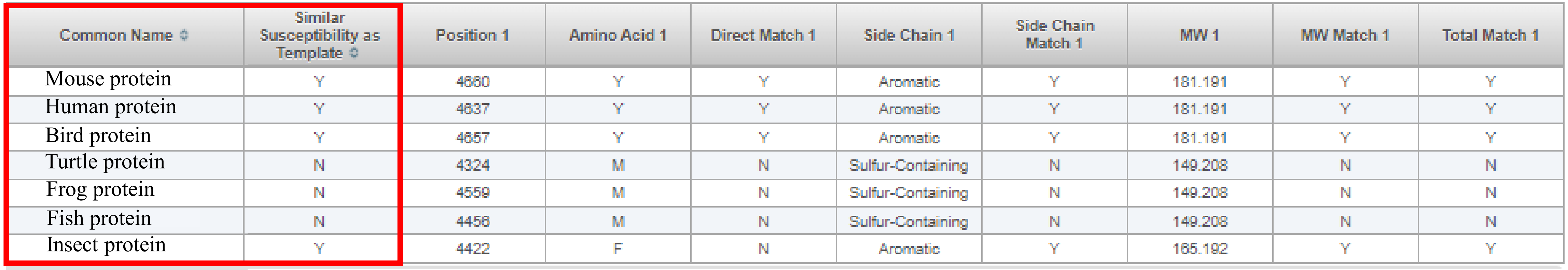
Full report:



Common Name	Similar Susceptibility as Template	Position 1	Amino Acid 1	Direct Match 1	Side Chain 1	Side Chain Match 1	MW 1	MW Match 1	Total Match 1
Mouse protein	Y	4660	Y	Y	Aromatic	Y	181.191	Y	Y
Human protein	Y	4637	Y	Y	Aromatic	Y	181.191	Y	Y
Bird protein	Y	4657	Y	Y	Aromatic	Y	181.191	Y	Y
Turtle protein	N	4324	M	N	Sulfur-Containing	N	149.208	N	N
Frog protein	N	4559	M	N	Sulfur-Containing	N	149.208	N	N
Fish protein	N	4456	M	N	Sulfur-Containing	N	149.208	N	N
Insect protein	Y	4422	F	N	Aromatic	Y	165.192	Y	Y

Level 3: Continued

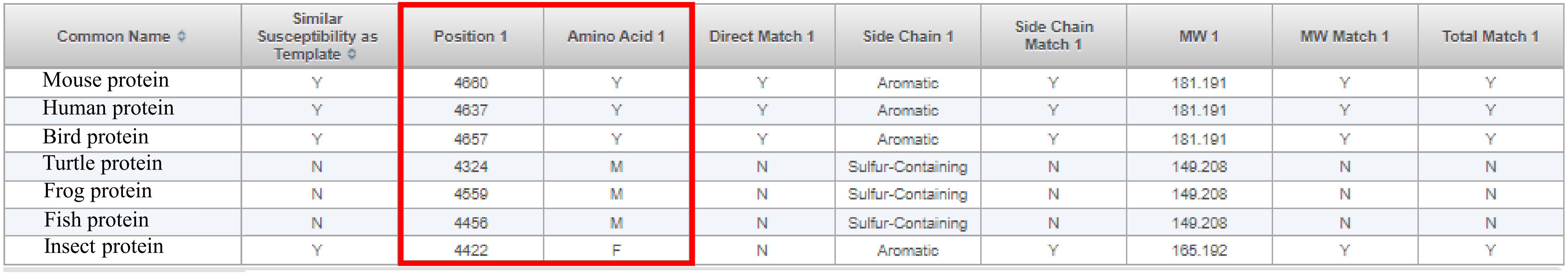
This example compares protein targets from 6 different species to Mouse as the query species.



Common Name	Similar Susceptibility as Template	Position 1	Amino Acid 1	Direct Match 1	Side Chain 1	Side Chain Match 1	MW 1	MW Match 1	Total Match 1
Mouse protein	Y	4660	Y	Y	Aromatic	Y	181.191	Y	Y
Human protein	Y	4637	Y	Y	Aromatic	Y	181.191	Y	Y
Bird protein	Y	4657	Y	Y	Aromatic	Y	181.191	Y	Y
Turtle protein	N	4324	M	N	Sulfur-Containing	N	149.208	N	N
Frog protein	N	4559	M	N	Sulfur-Containing	N	149.208	N	N
Fish protein	N	4456	M	N	Sulfur-Containing	N	149.208	N	N
Insect protein	Y	4422	F	N	Aromatic	Y	165.192	Y	Y

Level 3: Continued

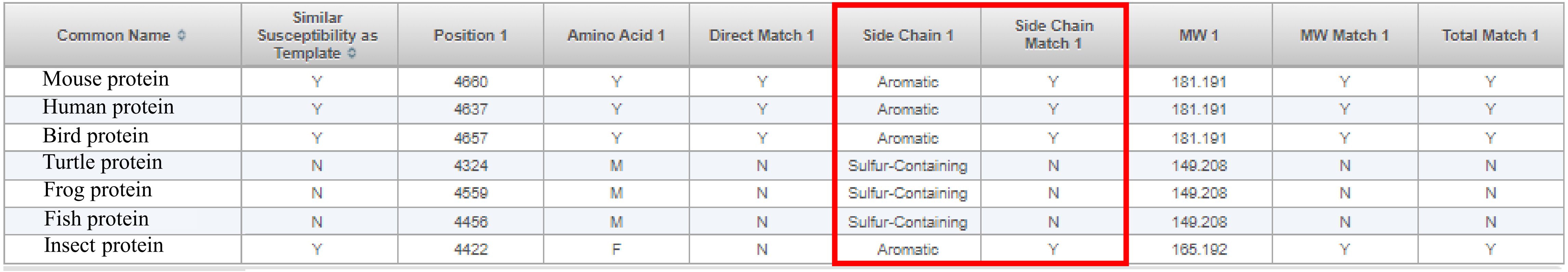
This example has 1 key amino acid position, but most proteins have multiple positions that are directly involved in protein-chemical or protein-protein interactions.



Common Name	Similar Susceptibility as Template	Position 1	Amino Acid 1	Direct Match 1	Side Chain 1	Side Chain Match 1	MW 1	MW Match 1	Total Match 1
Mouse protein	Y	4660	Y	Y	Aromatic	Y	181.191	Y	Y
Human protein	Y	4637	Y	Y	Aromatic	Y	181.191	Y	Y
Bird protein	Y	4657	Y	Y	Aromatic	Y	181.191	Y	Y
Turtle protein	N	4324	M	N	Sulfur-Containing	N	149.208	N	N
Frog protein	N	4559	M	N	Sulfur-Containing	N	149.208	N	N
Fish protein	N	4456	M	N	Sulfur-Containing	N	149.208	N	N
Insect protein	Y	4422	F	N	Aromatic	Y	165.192	Y	Y

Level 3: Continued

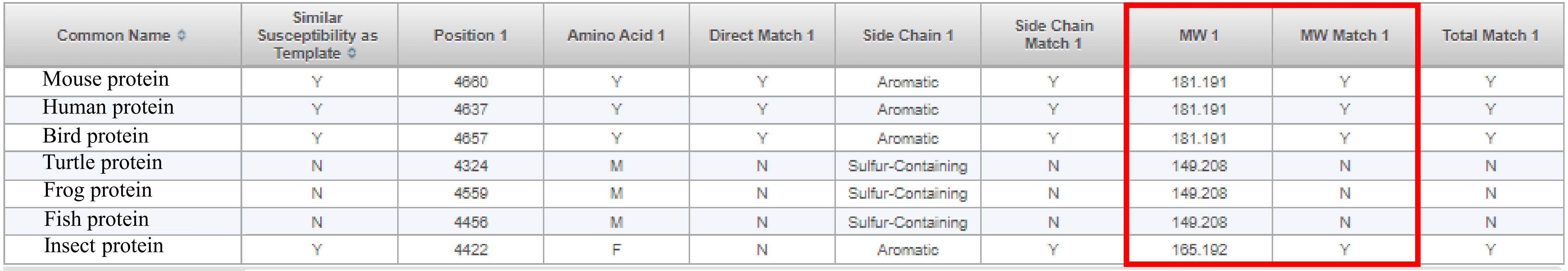
SeqAPASS compares the side-chain class of each amino acid from each species and whether it is a match to the query (mouse).



Common Name	Similar Susceptibility as Template	Position 1	Amino Acid 1	Direct Match 1	Side Chain 1	Side Chain Match 1	MW 1	MW Match 1	Total Match 1
Mouse protein	Y	4660	Y	Y	Aromatic	Y	181.191	Y	Y
Human protein	Y	4637	Y	Y	Aromatic	Y	181.191	Y	Y
Bird protein	Y	4657	Y	Y	Aromatic	Y	181.191	Y	Y
Turtle protein	N	4324	M	N	Sulfur-Containing	N	149.208	N	N
Frog protein	N	4559	M	N	Sulfur-Containing	N	149.208	N	N
Fish protein	N	4456	M	N	Sulfur-Containing	N	149.208	N	N
Insect protein	Y	4422	F	N	Aromatic	Y	165.192	Y	Y

Level 3: Continued

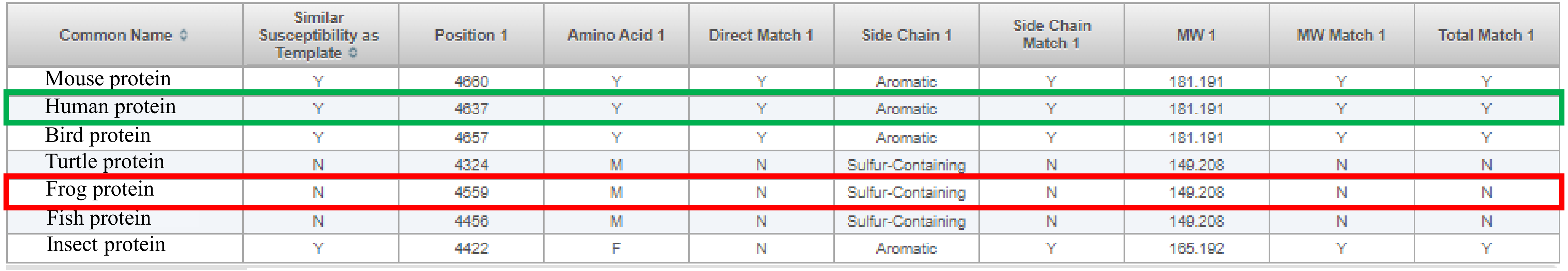
SeqAPASS compares the molecular weight of each amino acid from each species and whether the difference is less than 30 g/mol from query (mouse).



Common Name ⇕	Similar Susceptibility as Template ⇕	Position 1	Amino Acid 1	Direct Match 1	Side Chain 1	Side Chain Match 1	MW 1	MW Match 1	Total Match 1
Mouse protein	Y	4660	Y	Y	Aromatic	Y	181.191	Y	Y
Human protein	Y	4637	Y	Y	Aromatic	Y	181.191	Y	Y
Bird protein	Y	4657	Y	Y	Aromatic	Y	181.191	Y	Y
Turtle protein	N	4324	M	N	Sulfur-Containing	N	149.208	N	N
Frog protein	N	4559	M	N	Sulfur-Containing	N	149.208	N	N
Fish protein	N	4456	M	N	Sulfur-Containing	N	149.208	N	N
Insect protein	Y	4422	F	N	Aromatic	Y	165.192	Y	Y

Level 3: Continued

To provide conservative predictions, two “No” matches for one or more amino acids are required for SeqAPASS to predict that the species differ in susceptibility.

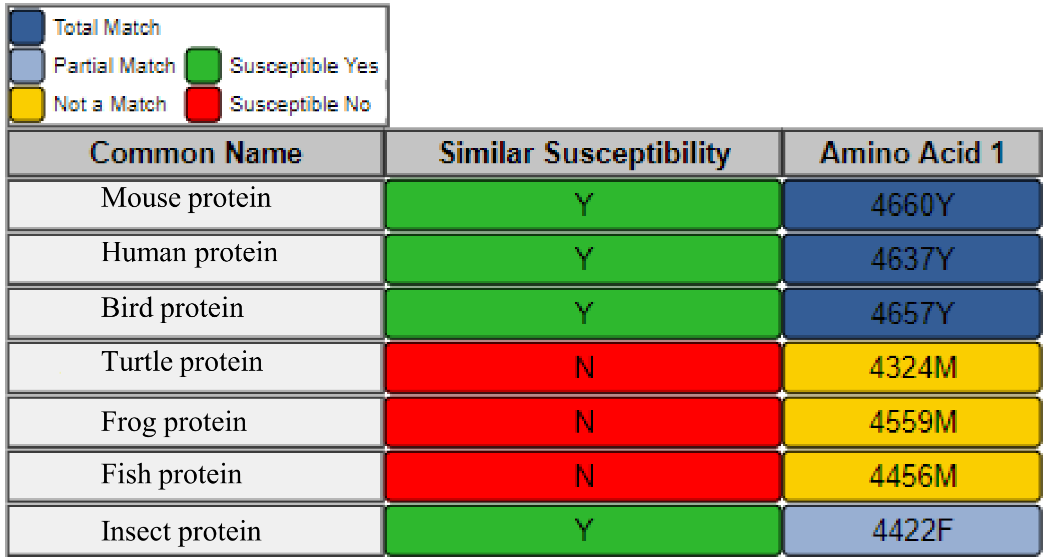


Common Name ⇕	Similar Susceptibility as Template ⇕	Position 1	Amino Acid 1	Direct Match 1	Side Chain 1	Side Chain Match 1	MW 1	MW Match 1	Total Match 1
Mouse protein	Y	4660	Y	Y	Aromatic	Y	181.191	Y	Y
Human protein	Y	4637	Y	Y	Aromatic	Y	181.191	Y	Y
Bird protein	Y	4657	Y	Y	Aromatic	Y	181.191	Y	Y
Turtle protein	N	4324	M	N	Sulfur-Containing	N	149.208	N	N
Frog protein	N	4559	M	N	Sulfur-Containing	N	149.208	N	N
Fish protein	N	4456	M	N	Sulfur-Containing	N	149.208	N	N
Insect protein	Y	4422	F	N	Aromatic	Y	165.192	Y	Y

Level 3: Visualization

Heat Map Visualization!

Simple report:



🟦 Total Match
▢ Partial Match 🟩 Susceptible Yes
🟨 Not a Match 🟥 Susceptible No

Common Name	Similar Susceptibility	Amino Acid 1
Mouse protein	Y	4660Y
Human protein	Y	4637Y
Bird protein	Y	4657Y
Turtle protein	N	4324M
Frog protein	N	4559M
Fish protein	N	4456M
Insect protein	Y	4422F

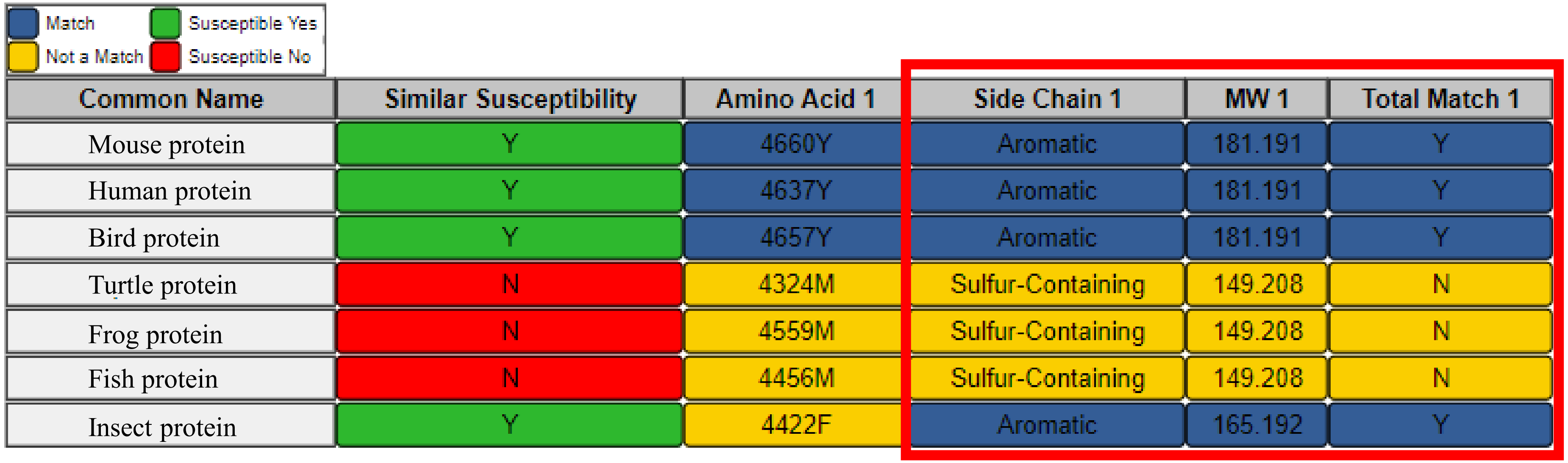
Customizable:

- Common/scientific name
- Ortholog candidates
- Endangered species
- Threatened species
- Common model organisms
- Amino acid information displayed

Level 3: Visualization

Heat Map Visualization!

Full report:

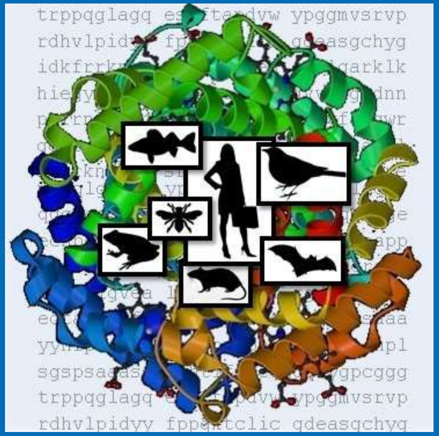


🟦 Match 🟩 Susceptible Yes
🟨 Not a Match 🟥 Susceptible No

Common Name	Similar Susceptibility	Amino Acid 1	Side Chain 1	MW 1	Total Match 1
Mouse protein	Y	4660Y	Aromatic	181.191	Y
Human protein	Y	4637Y	Aromatic	181.191	Y
Bird protein	Y	4657Y	Aromatic	181.191	Y
Turtle protein	N	4324M	Sulfur-Containing	149.208	N
Frog protein	N	4559M	Sulfur-Containing	149.208	N
Fish protein	N	4456M	Sulfur-Containing	149.208	N
Insect protein	Y	4422F	Aromatic	165.192	Y

SeqAPASS Live Demo

Level 3 setup, output, and visualization



United States
Environmental Protection
Agency

Training Course - SeqAPASS v7.1: Decision Summary Report

Instructors:

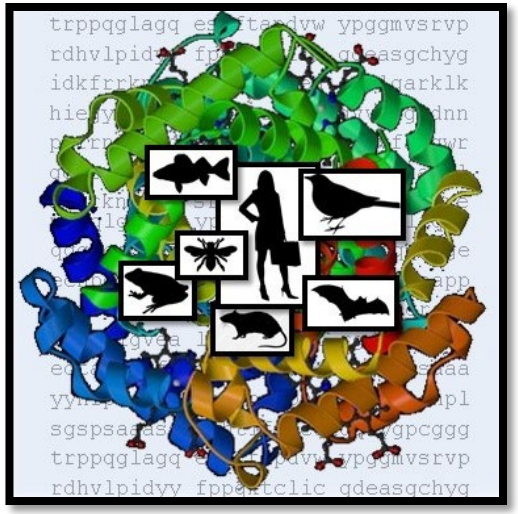
Carlie A. LaLone, Ph.D. (US EPA)

Marissa Brickley, PhD student (UMN-Duluth/ US EPA)

Office of Research and Development
Center for Computational Toxicology and Exposure, Great Lakes Toxicology and Ecology Division

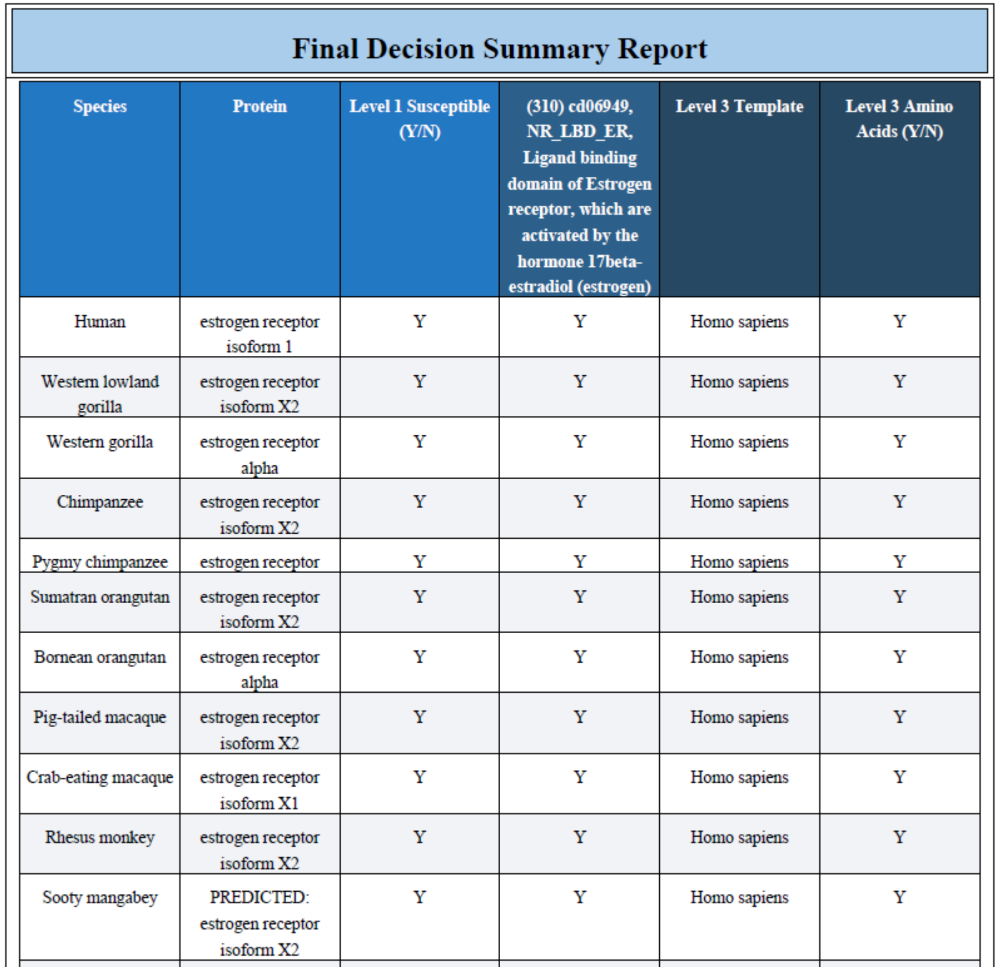
Training Course Overview

- **Introduction** to challenges in species extrapolation and how bioinformatics can aid in addressing these challenges (LaLone)
- Training with SeqAPASS
  - Level 1 and Level 1 Data visualization (LaLone)
  - Level 2 and Level 2 Data visualization (Jensen)
    - Instructor Demos – Level 1 and 2
  - Level 3 and Level 3 Data visualization (Jensen)
    - Instructor Demos – Level 3
  - Decision summary report (Jensen)
  - Level 4 (LaLone)



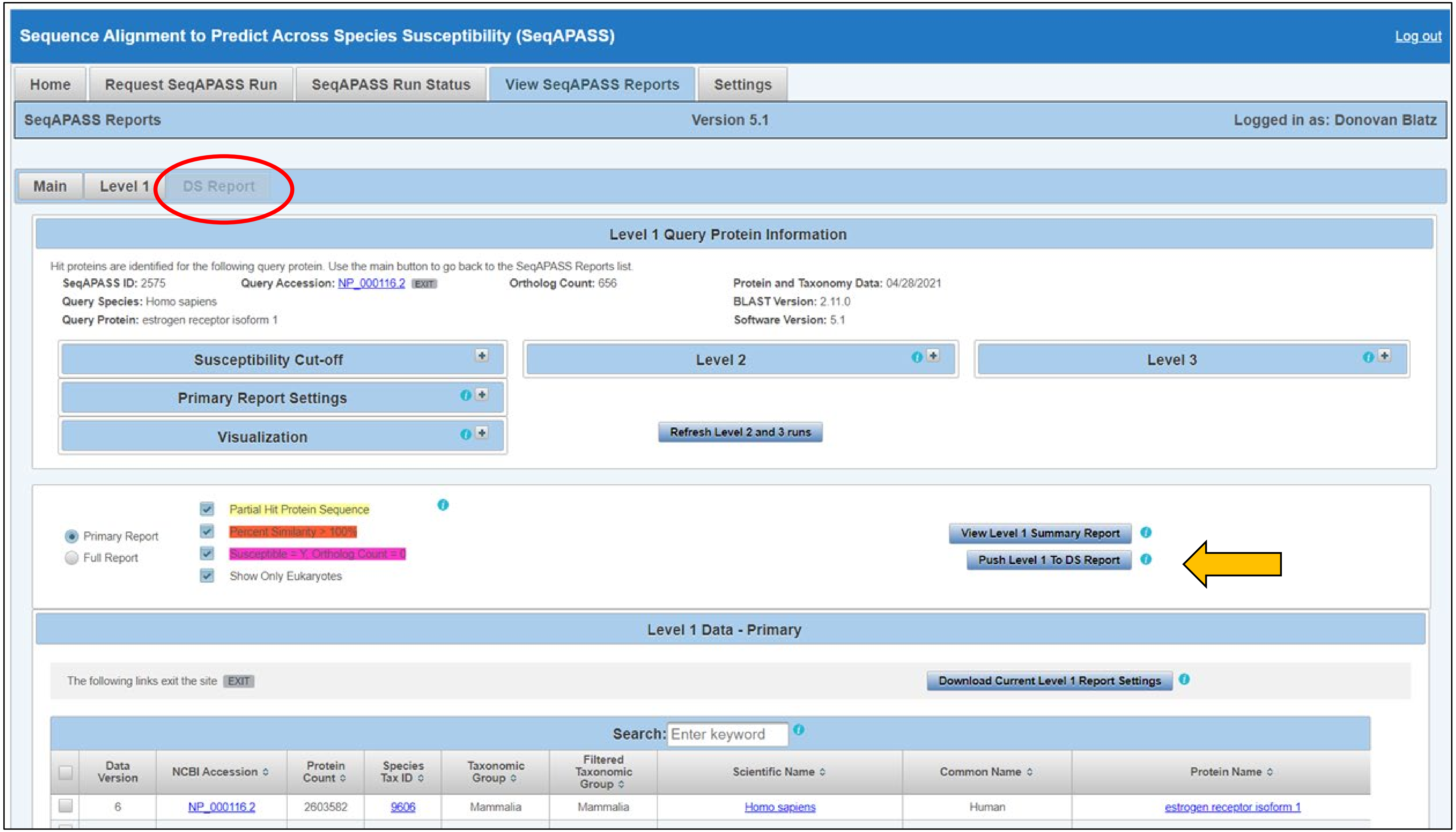
Uses for the Decision Summary (DS) Report

- A great way to organize data from level runs
  - Can contain one Level 1, multiple Level 2, and one Level 3 run(s)
- Can be customized to display the desired data into a downloadable PDF
- Provides manuscript ready visualizations



Final Decision Summary Report

Species	Protein	Level 1 Susceptible (Y/N)	(310) cd06949, NR_LBD_ER, Ligand binding domain of Estrogen receptor, which are activated by the hormone 17beta-estradiol (estrogen)	Level 3 Template	Level 3 Amino Acids (Y/N)
Human	estrogen receptor isoform 1	Y	Y	Homo sapiens	Y
Western lowland gorilla	estrogen receptor isoform X2	Y	Y	Homo sapiens	Y
Western gorilla	estrogen receptor alpha	Y	Y	Homo sapiens	Y
Chimpanzee	estrogen receptor isoform X2	Y	Y	Homo sapiens	Y
Pygmy chimpanzee	estrogen receptor	Y	Y	Homo sapiens	Y
Sumatran orangutan	estrogen receptor isoform X2	Y	Y	Homo sapiens	Y
Bornean orangutan	estrogen receptor alpha	Y	Y	Homo sapiens	Y
Pig-tailed macaque	estrogen receptor isoform X2	Y	Y	Homo sapiens	Y
Crab-eating macaque	estrogen receptor isoform X1	Y	Y	Homo sapiens	Y
Rhesus monkey	estrogen receptor isoform X2	Y	Y	Homo sapiens	Y
Sooty mangabey	PREDICTED: estrogen receptor isoform X2	Y	Y	Homo sapiens	Y



Hit proteins are identified for the following query protein. Use the main button to go back to the SeqAPASS Reports list.

SeqAPASS ID: 2575
Query Accession: NP_000116.2 EXIT
Ortholog Count: 656
Protein and Taxonomy Data: 04/28/2021
Query Species: Homo sapiens
BLAST Version: 2.11.0
Query Protein: estrogen receptor isoform 1
Software Version: 5.1

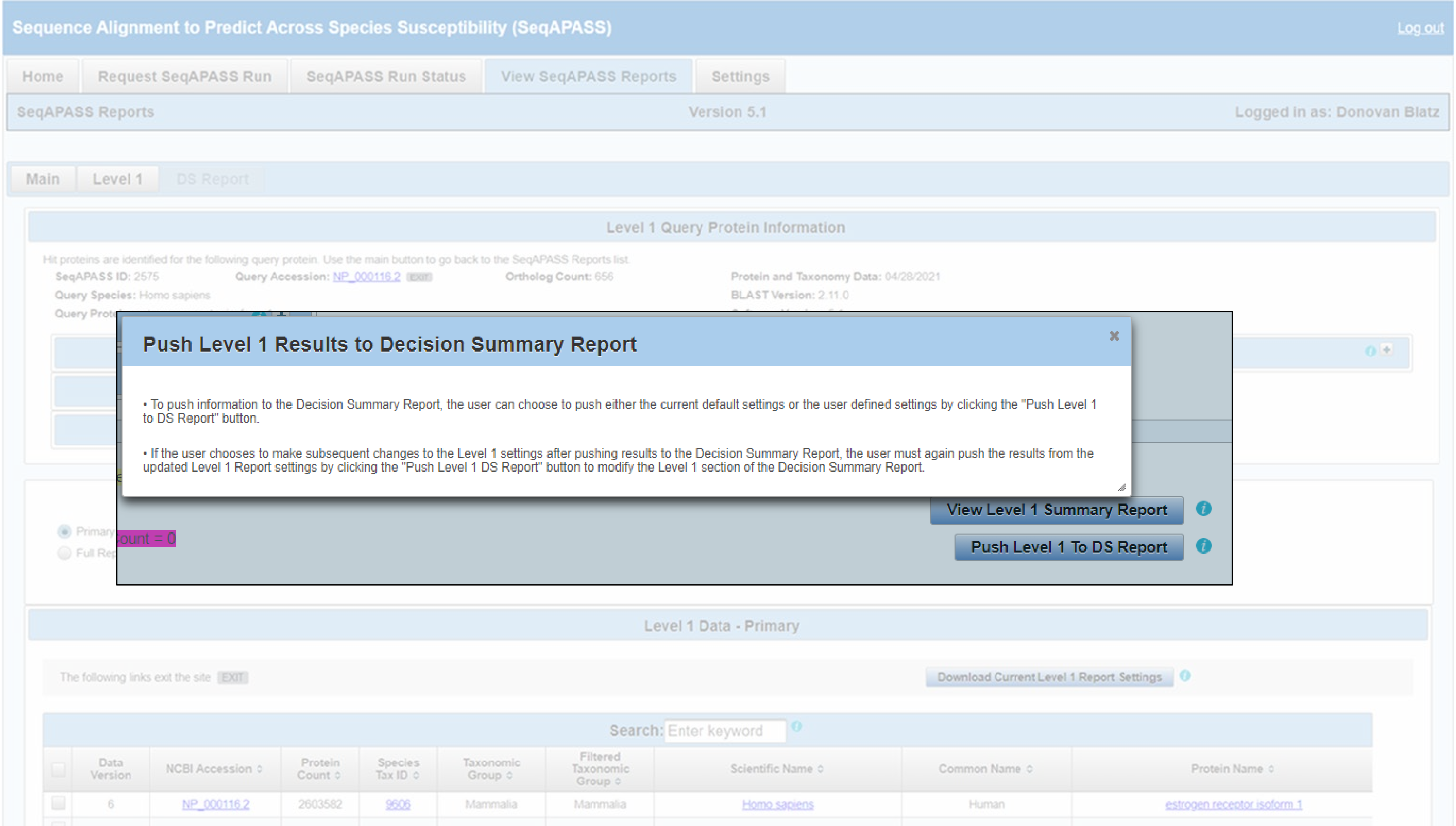
● Primary Report
○ Full Report

☑ Partial Hit Protein Sequence
☑ Percent Similarity > 100%
☑ Susceptible = Y, Ortholog Count = 0
☑ Show Only Eukaryotes

The following links exit the site EXIT

Search: Enter keyword

☐	Data Version	NCBI Accession ⇕	Protein Count ⇕	Species Tax ID ⇕	Taxonomic Group ⇕	Filtered Taxonomic Group ⇕	Scientific Name ⇕	Common Name ⇕	Protein Name ⇕
☐	6	NP_000116.2	2603582	9606	Mammalia	Mammalia	Homo sapiens	Human	estrogen receptor isoform 1



Hit proteins are identified for the following query protein. Use the main button to go back to the SeqAPASS Reports list.

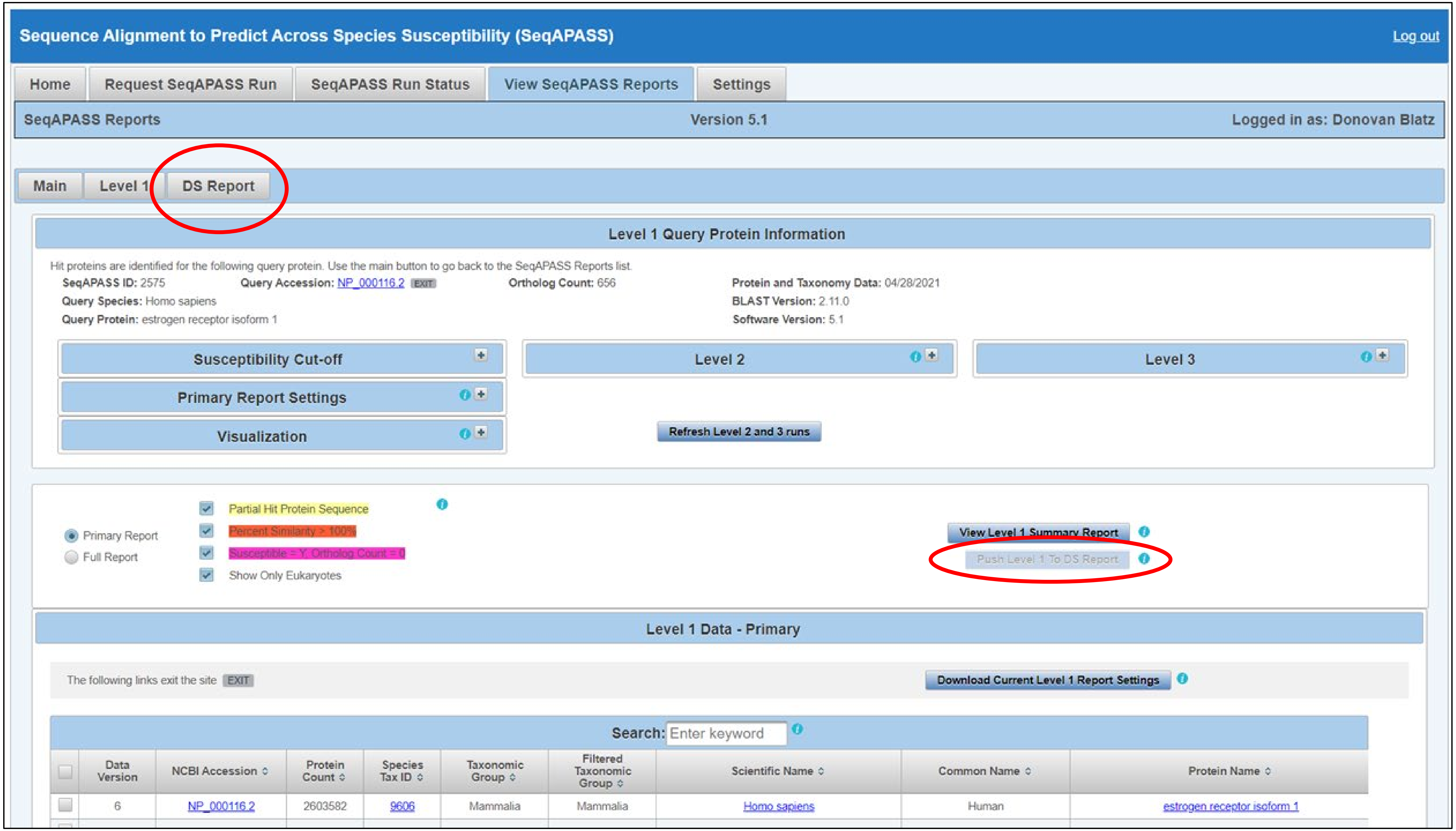
SeqAPASS ID: 2575
Query Accession: NP_000116.2 EXIT
Ortholog Count: 656
Protein and Taxonomy Data: 04/28/2021
Query Species: Homo sapiens
BLAST Version: 2.11.0

Push Level 1 Results to Decision Summary Report

- To push information to the Decision Summary Report, the user can choose to push either the current default settings or the user defined settings by clicking the "Push Level 1 to DS Report" button.

- If the user chooses to make subsequent changes to the Level 1 settings after pushing results to the Decision Summary Report, the user must again push the results from the updated Level 1 Report settings by clicking the "Push Level 1 DS Report" button to modify the Level 1 section of the Decision Summary Report.

☐	Data Version	NCBI Accession	Protein Count	Species Tax ID	Taxonomic Group	Filtered Taxonomic Group	Scientific Name	Common Name	Protein Name
☐	6	NP_000116.2	2603582	9606	Mammalia	Mammalia	Homo sapiens	Human	estrogen receptor isoform 1



Hit proteins are identified for the following query protein. Use the main button to go back to the SeqAPASS Reports list.

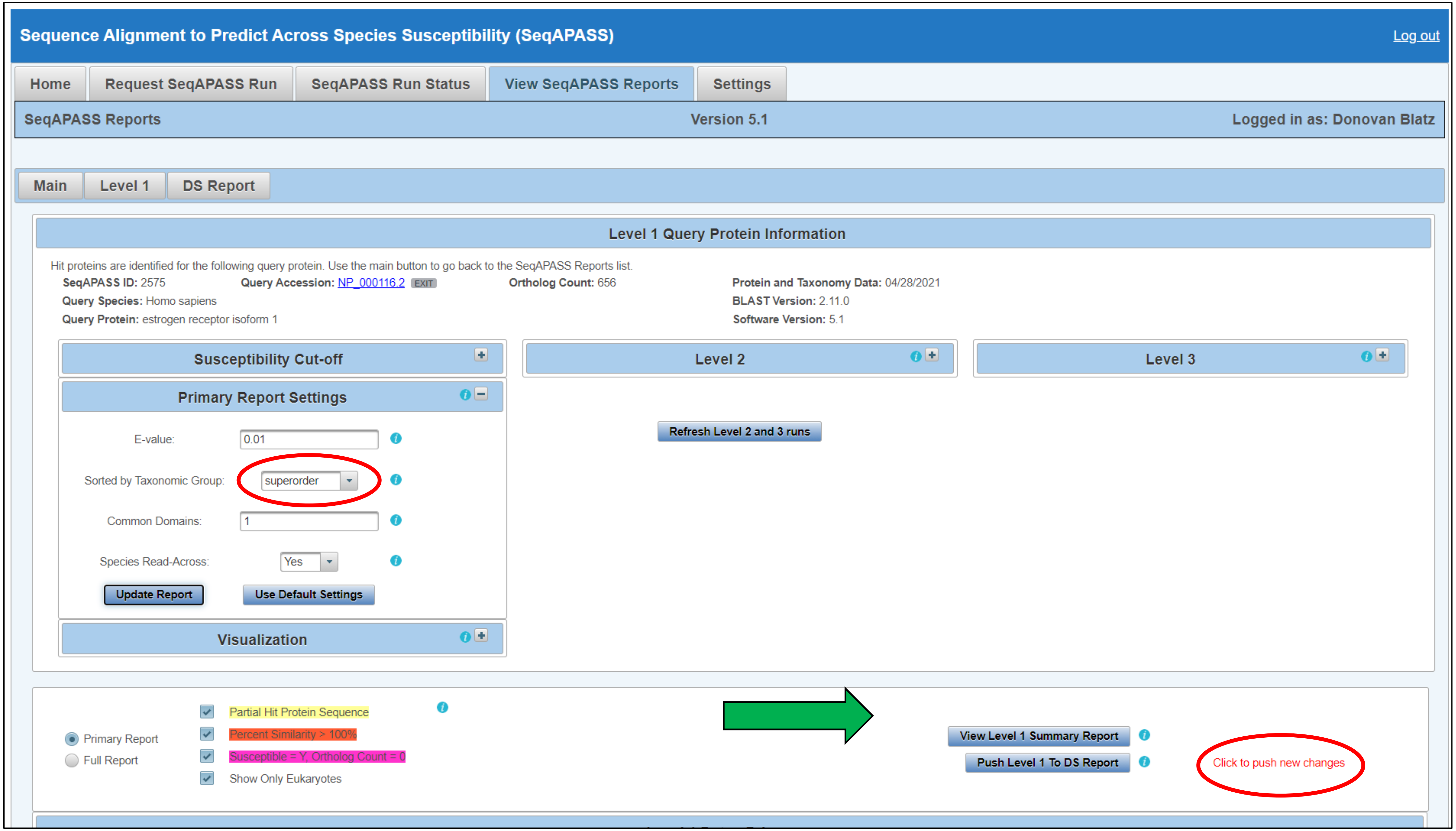
SeqAPASS ID: 2575
Query Accession: NP_000116.2 EXIT
Ortholog Count: 656
Protein and Taxonomy Data: 04/28/2021
Query Species: Homo sapiens
BLAST Version: 2.11.0
Query Protein: estrogen receptor isoform 1
Software Version: 5.1

● Primary Report
○ Full Report

☑ Partial Hit Protein Sequence
☑ Percent Similarity > 100%
☑ Susceptible = Y, Ortholog Count = 0
☑ Show Only Eukaryotes

The following links exit the site EXIT

☐	Data Version	NCBI Accession	Protein Count	Species Tax ID	Taxonomic Group	Filtered Taxonomic Group	Scientific Name	Common Name	Protein Name
☐	6	NP_000116.2	2603582	9606	Mammalia	Mammalia	Homo sapiens	Human	estrogen receptor isoform 1



Level 1 Query Protein Information

Hit proteins are identified for the following query protein. Use the main button to go back to the SeqAPASS Reports list.

SeqAPASS ID: 2575
Query Accession: NP_000116.2 EXIT
Ortholog Count: 656
Protein and Taxonomy Data: 04/28/2021
Query Species: Homo sapiens
BLAST Version: 2.11.0
Query Protein: estrogen receptor isoform 1
Software Version: 5.1

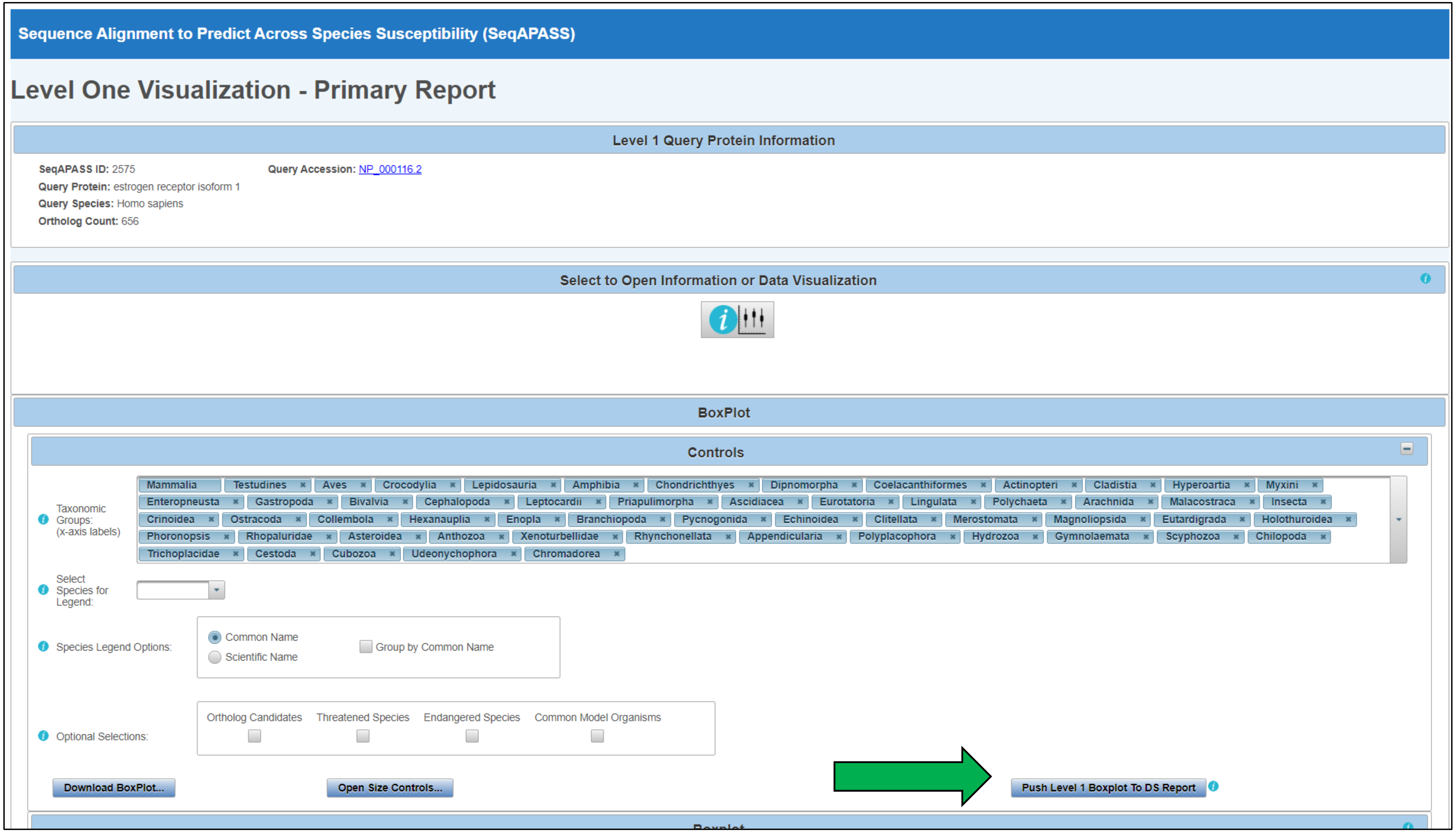
Primary Report Settings

E-value: 0.01
Sorted by Taxonomic Group: superorder
Common Domains: 1
Species Read-Across: Yes

● Primary Report
○ Full Report

☑ Partial Hit Protein Sequence
☑ Percent Similarity > 100%
☑ Susceptible = Y, Ortholog Count = 0
☑ Show Only Eukaryotes

Click to push new changes



Sequence Alignment to Predict Across Species Susceptibility (SeqAPASS)

Level One Visualization - Primary Report

Level 1 Query Protein Information

SeqAPASS ID: 2575
Query Protein: estrogen receptor isoform 1
Query Species: Homo sapiens
Ortholog Count: 656
Query Accession: NP_000116.2

Select to Open Information or Data Visualization

BoxPlot

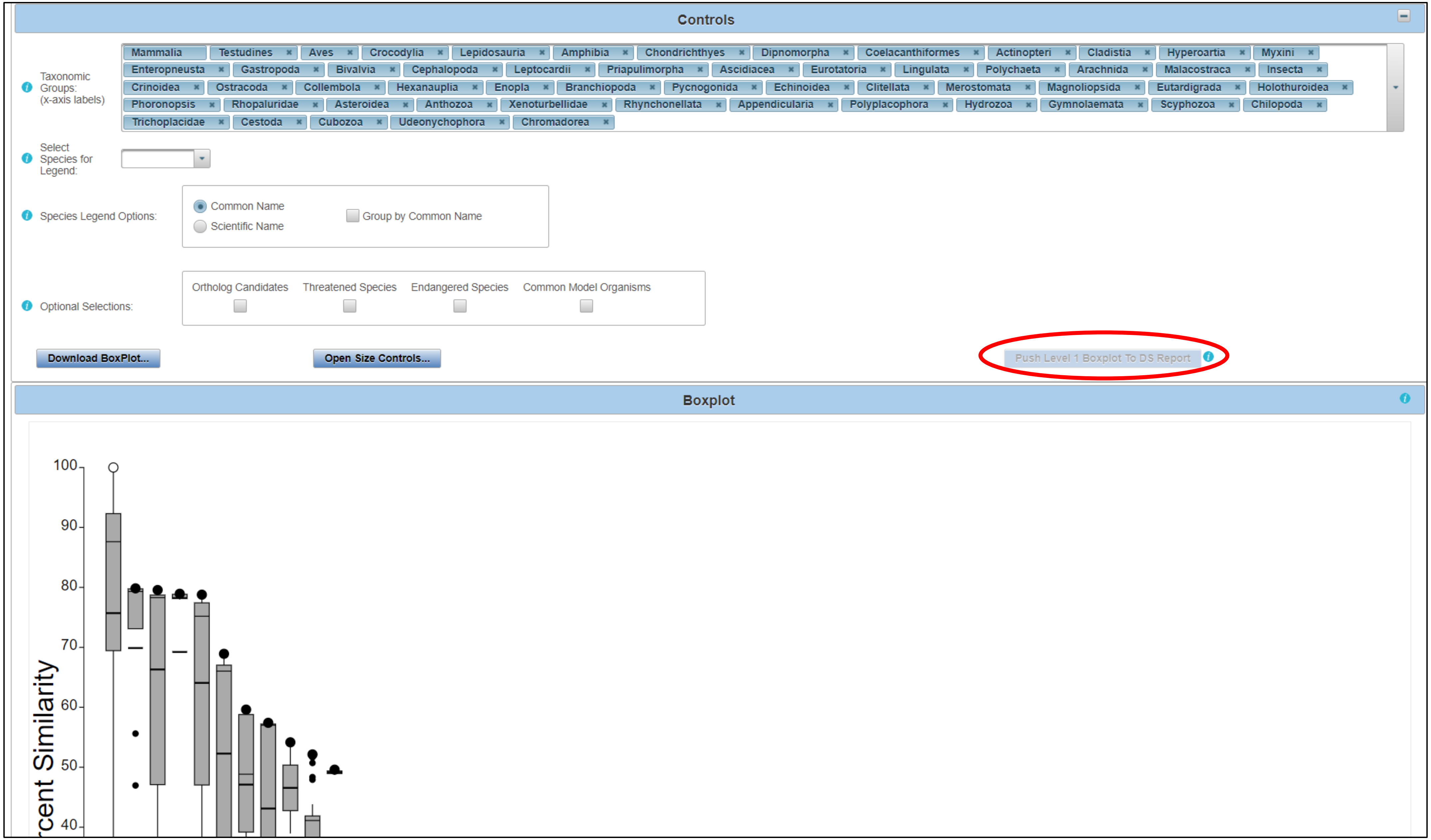
Controls

Taxonomic Groups: (x-axis labels)
Mammalia, Testudines, Aves, Crocodylia, Lepidosauria, Amphibia, Chondrichthyes, Dipnomorpha, Coelacanthiformes, Actinopteri, Cladistia, Hyperoartia, Myxini
Enteropneusta, Gastropoda, Bivalvia, Cephalopoda, Leptocardii, Priapulimorpha, Ascidiacea, Eurotatoria, Lingulata, Polychaeta, Arachnida, Malacostraca, Insecta
Crinoidea, Ostracoda, Collembola, Hexanauplia, Enopla, Branchiopoda, Pycnogonida, Echinoidea, Clitellata, Merostomata, Magnoliopsida, Eutardigrada, Holothuroidea
Phoronopsis, Rhopaluridae, Asteroidea, Anthozoa, Xenoturbellidae, Rhynchonellata, Appendicularia, Polyplacophora, Hydrozoa, Gymnolaemata, Scyphozoa, Chilopoda
Trichoplacidae, Cestoda, Cubozoa, Udeonychophora, Chromadorea

Select Species for Legend:

Species Legend Options: ● Common Name ○ Scientific Name ☐ Group by Common Name

Optional Selections: ☐ Ortholog Candidates ☐ Threatened Species ☐ Endangered Species ☐ Common Model Organisms



Controls

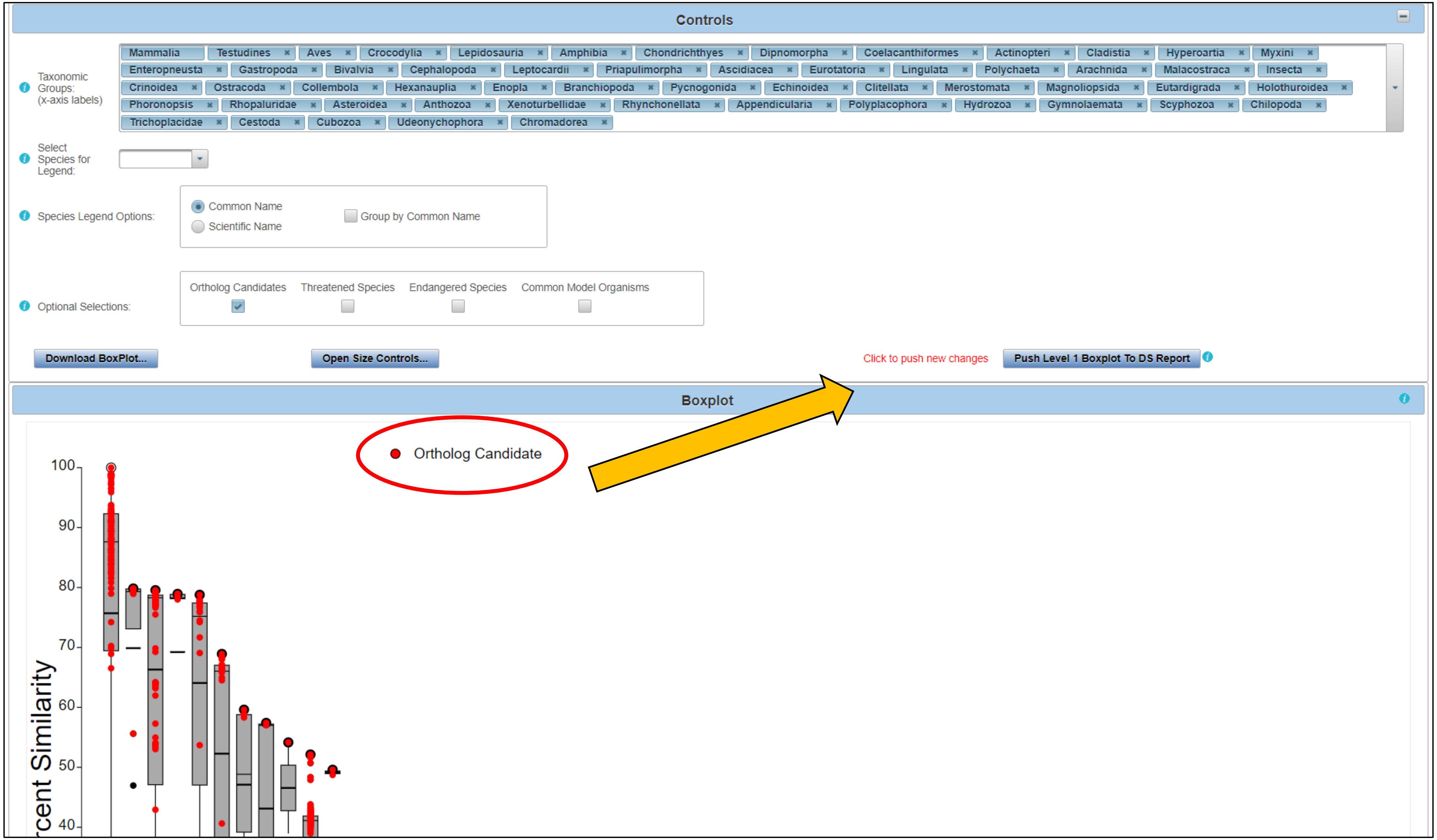
Taxonomic Groups: (x-axis labels)
Mammalia, Testudines, Aves, Crocodylia, Lepidosauria, Amphibia, Chondrichthyes, Dipnomorpha, Coelacanthiformes, Actinopteri, Cladistia, Hyperoartia, Myxini, Enteropneusta, Gastropoda, Bivalvia, Cephalopoda, Leptocardii, Priapulimorpha, Ascidiacea, Eurotatoria, Lingulata, Polychaeta, Arachnida, Malacostraca, Insecta, Crinoidea, Ostracoda, Collembola, Hexanauplia, Enopla, Branchiopoda, Pycnogonida, Echinoidea, Clitellata, Merostomata, Magnoliopsida, Eutardigrada, Holothuroidea, Phoronopsis, Rhopaluridae, Asteroidea, Anthozoa, Xenoturbellidae, Rhynchonellata, Appendicularia, Polyplacophora, Hydrozoa, Gymnolaemata, Scyphozoa, Chilopoda, Trichoplacidae, Cestoda, Cubozoa, Udeonychophora, Chromadorea

Select Species for Legend:

Species Legend Options: ● Common Name ○ Scientific Name ☐ Group by Common Name

Optional Selections: Ortholog Candidates ☐ Threatened Species ☐ Endangered Species ☐ Common Model Organisms ☐

Boxplot



Controls

Taxonomic Groups: (x-axis labels)
Mammalia, Testudines, Aves, Crocodylia, Lepidosauria, Amphibia, Chondrichthyes, Dipnomorpha, Coelacanthiformes, Actinopteri, Cladistia, Hyperoartia, Myxini, Enteropneusta, Gastropoda, Bivalvia, Cephalopoda, Leptocardii, Priapulimorpha, Ascidiacea, Eurotatoria, Lingulata, Polychaeta, Arachnida, Malacostraca, Insecta, Crinoidea, Ostracoda, Collembola, Hexanauplia, Enopla, Branchiopoda, Pycnogonida, Echinoidea, Clitellata, Merostomata, Magnoliopsida, Eutardigrada, Holothuroidea, Phoronopsis, Rhopaluridae, Asteroidea, Anthozoa, Xenoturbellidae, Rhynchonellata, Appendicularia, Polyplacophora, Hydrozoa, Gymnolaemata, Scyphozoa, Chilopoda, Trichoplacidae, Cestoda, Cubozoa, Udeonychophora, Chromadorea

Select Species for Legend:

Species Legend Options:
● Common Name
○ Scientific Name
☐ Group by Common Name

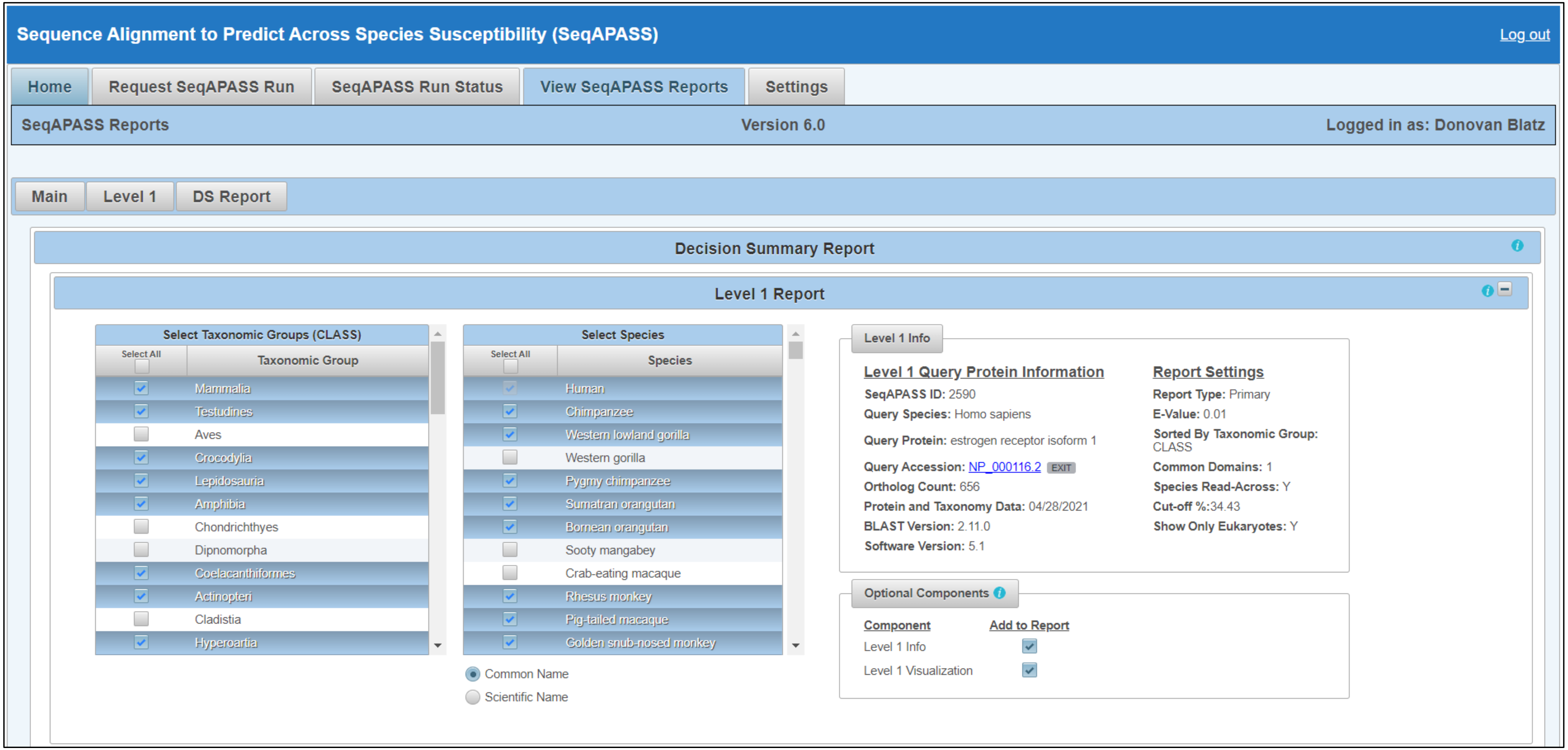
Optional Selections:
☑ Ortholog Candidates
☐ Threatened Species
☐ Endangered Species
☐ Common Model Organisms

Click to push new changes

Boxplot

● Ortholog Candidate

Time for hands-on training with the DS report



Select Taxonomic Groups (CLASS)

Select All ☐	Taxonomic Group
☑	Mammalia
☑	Testudines
☐	Aves
☑	Crocodylia
☑	Lepidosauria
☑	Amphibia
☐	Chondrichthyes
☐	Dipnomorpha
☑	Coelacanthiformes
☑	Actinopteri
☐	Cladistia
☑	Hyperoartia

Select Species

Select All ☐	Species
☑	Human
☑	Chimpanzee
☑	Western lowland gorilla
☐	Western gorilla
☑	Pygmy chimpanzee
☑	Sumatran orangutan
☑	Bornean orangutan
☐	Sooty mangabey
☐	Crab-eating macaque
☑	Rhesus monkey
☑	Pig-tailed macaque
☑	Golden snub-nosed monkey

● Common Name
○ Scientific Name

Level 1 Info

Level 1 Query Protein Information

SeqAPASS ID: 2590
Query Species: Homo sapiens
Query Protein: estrogen receptor isoform 1
Query Accession: NP_000116.2 EXIT
Ortholog Count: 656
Protein and Taxonomy Data: 04/28/2021
BLAST Version: 2.11.0
Software Version: 5.1

Report Settings

Report Type: Primary
E-Value: 0.01
Sorted By Taxonomic Group: CLASS
Common Domains: 1
Species Read-Across: Y
Cut-off %:34.43
Show Only Eukaryotes: Y

Optional Components

Component	Add to Report
Level 1 Info	☑
Level 1 Visualization	☑

United States Environmental Protection Agency

Training Course – SeqAPASS v7.1 Level 4: Protein Structural Comparisons Across Species

Instructors:

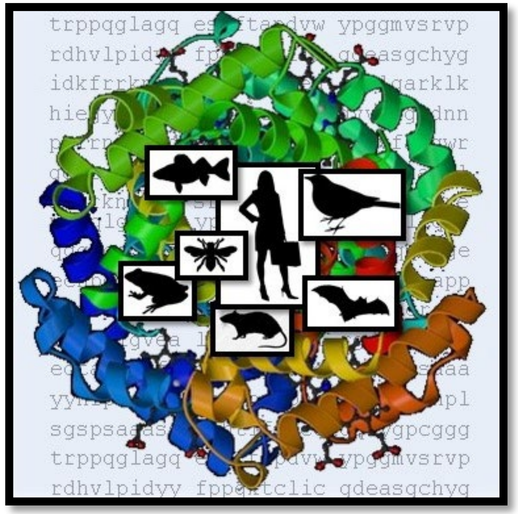
Carlie A. LaLone, Ph.D. (US EPA)

Marissa Brickley, PhD student (UMN-Duluth/ US EPA)

Office of Research and Development
Center for Computational Toxicology and Exposure, Great Lakes Toxicology and Ecology Division

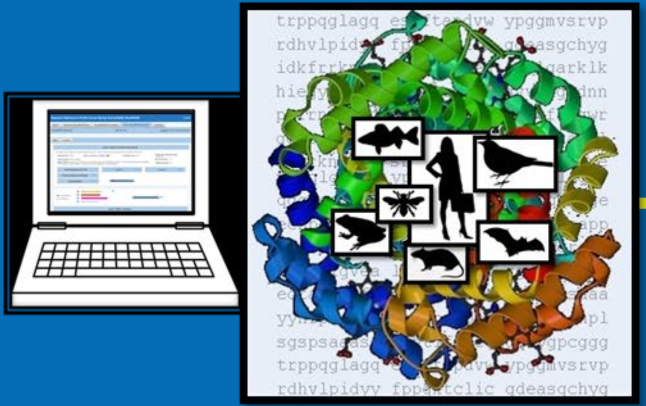
Training Course Overview

- **Introduction** to challenges in species extrapolation and how bioinformatics can aid in addressing these challenges (LaLone)
- Training with SeqAPASS
  - Level 1 and Level 1 Data visualization (LaLone)
  - Level 2 and Level 2 Data visualization (Jensen)
    - Instructor Demos – Level 1 and 2
  - Level 3 and Level 3 Data visualization (Jensen)
    - Instructor Demos – Level 3
  - Decision summary report (Jensen)
  - Level 4 (LaLone)

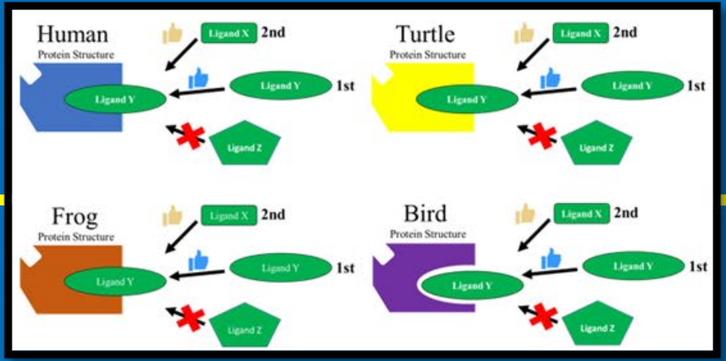


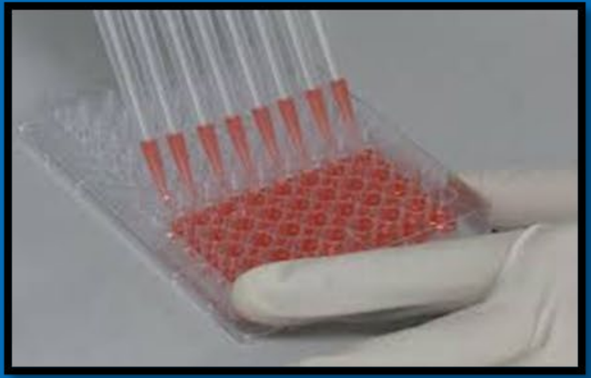
Begin Simple and Advance as the Science Advances

SeqAPASS v6

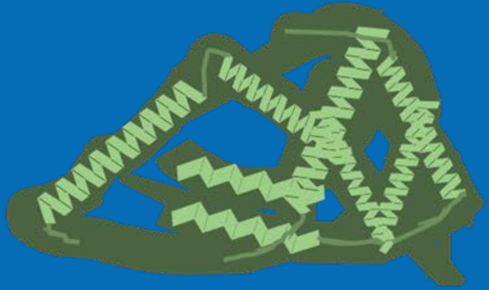


SeqAPASS v7





Sequence



Structure

Function

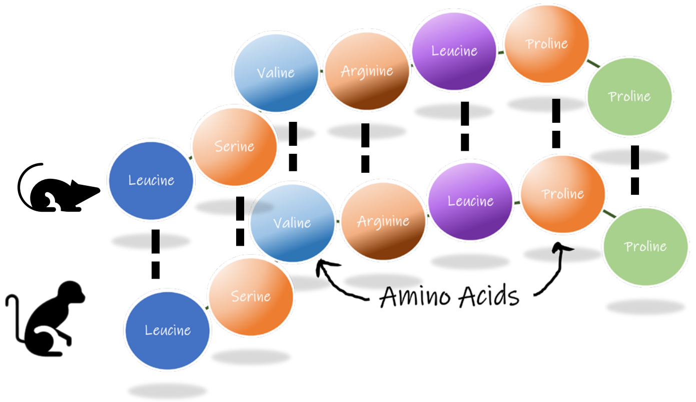
Consider sequence and structural attributes to understand protein conservation across species

Flexible Analysis Based On Available Data

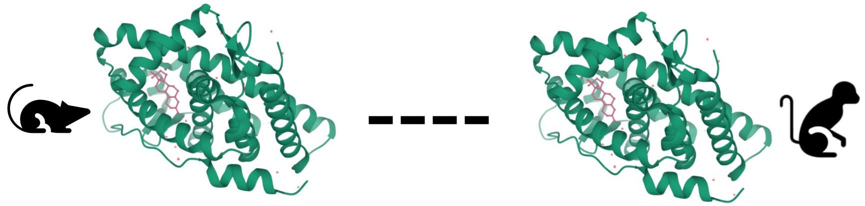
Level 1 Primary Amino Acid Sequence Alignments

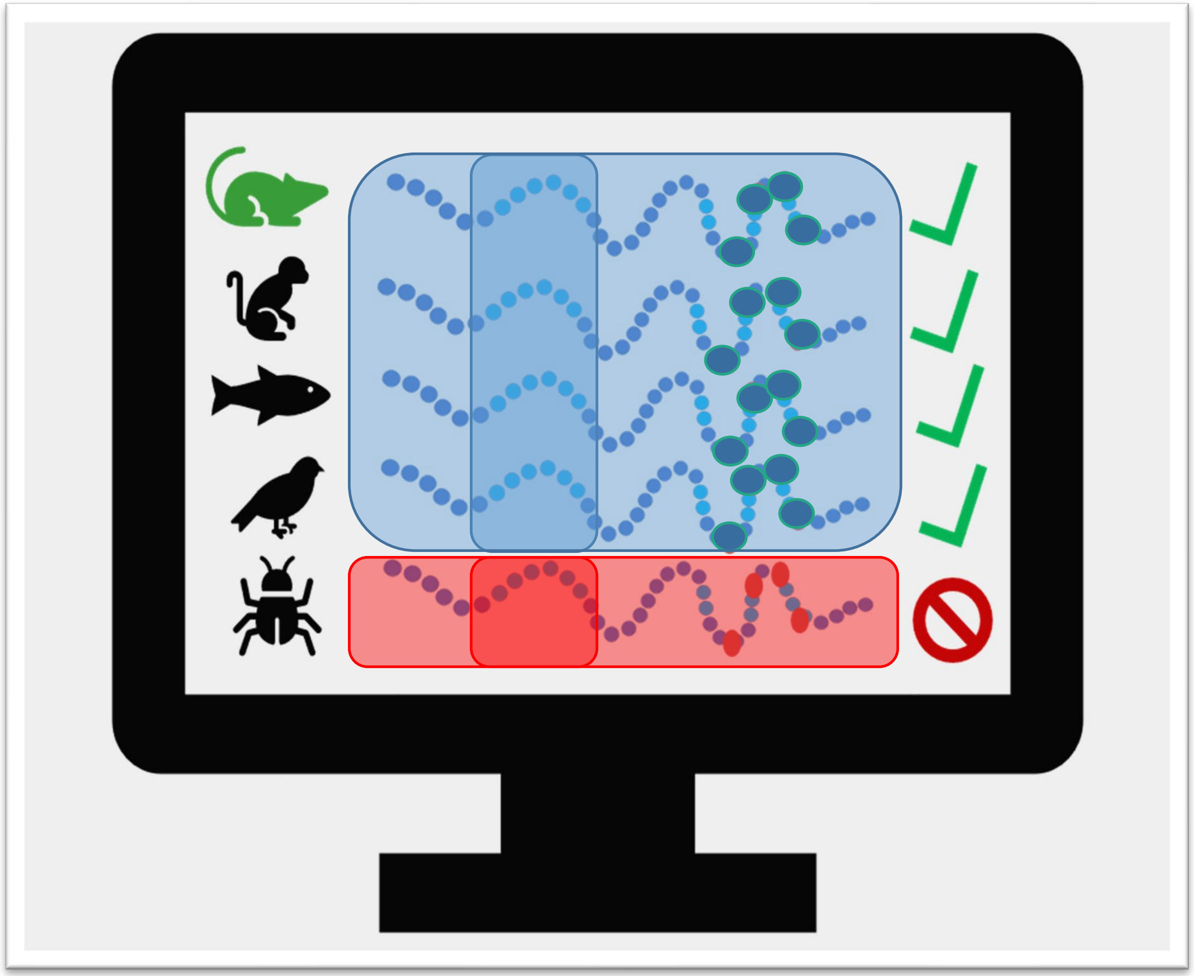
Level 2 Conserved Functional Domain Alignments

Level 3 Critical (Close Contact) Amino Acid Conservation

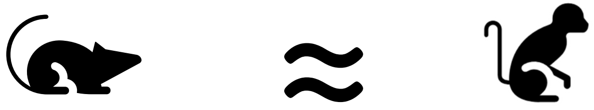


SeqAPASS v7.0 → Level 4 (expert users only) Structure Alignments

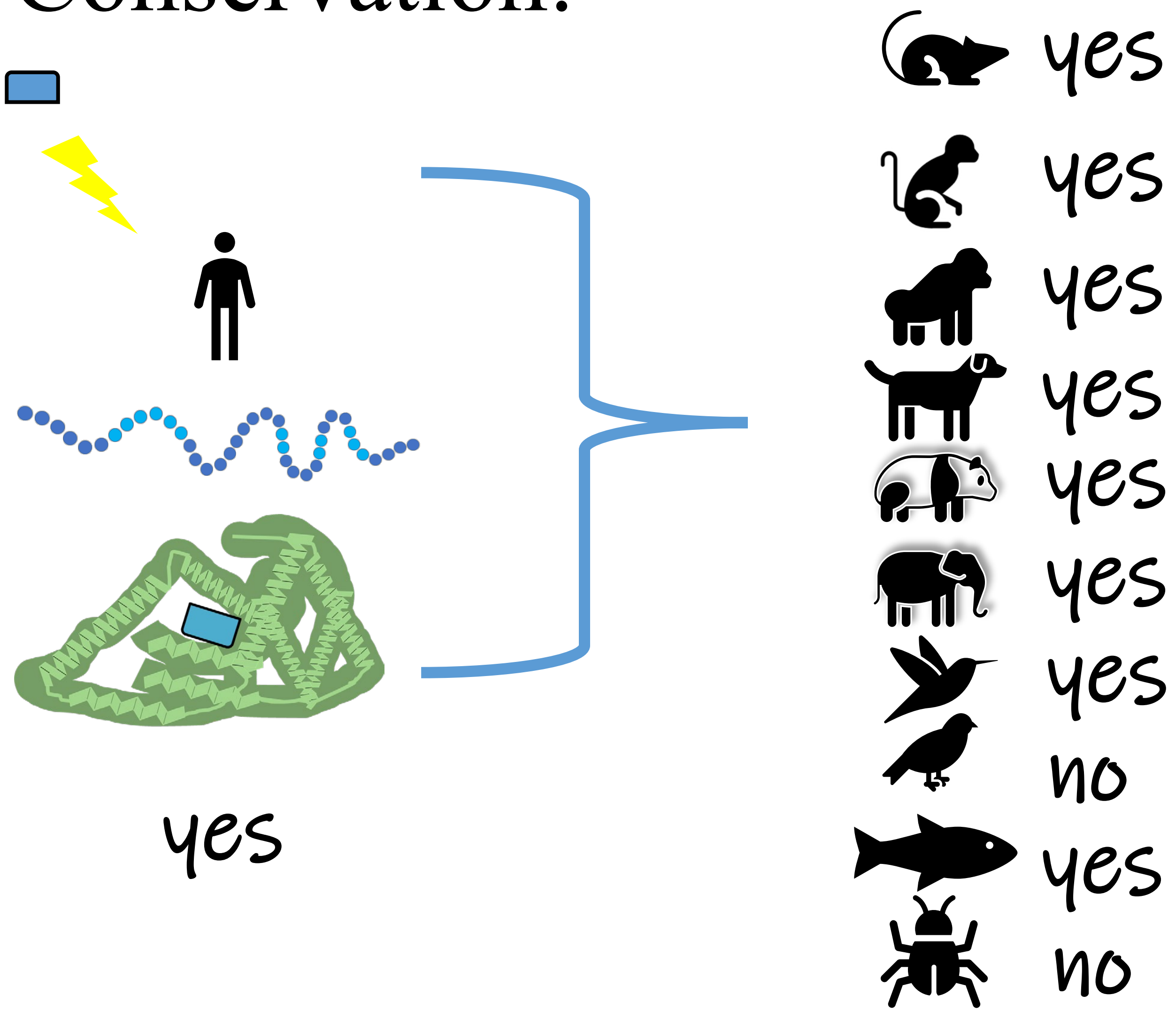




Gather Lines of Evidence Toward Protein Conservation

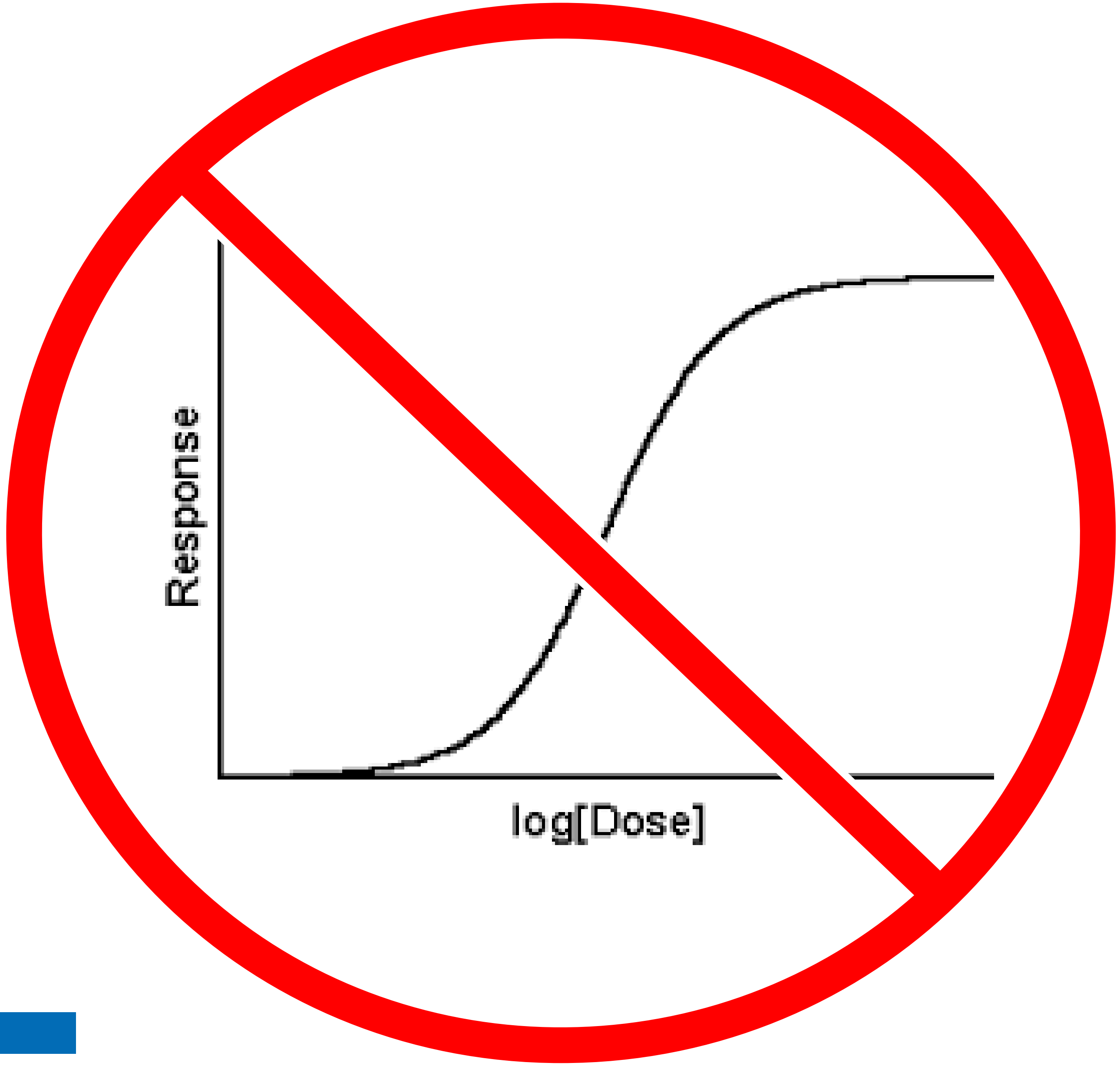


SeqAPASS Predicts Likelihood of Similar Susceptibility based on Sequence Conservation:



Line(s) of evidence indicate
- The protein is conserved
- The protein is NOT conserved

SeqAPASS DOES NOT predict the degree of sensitivity/susceptibility:



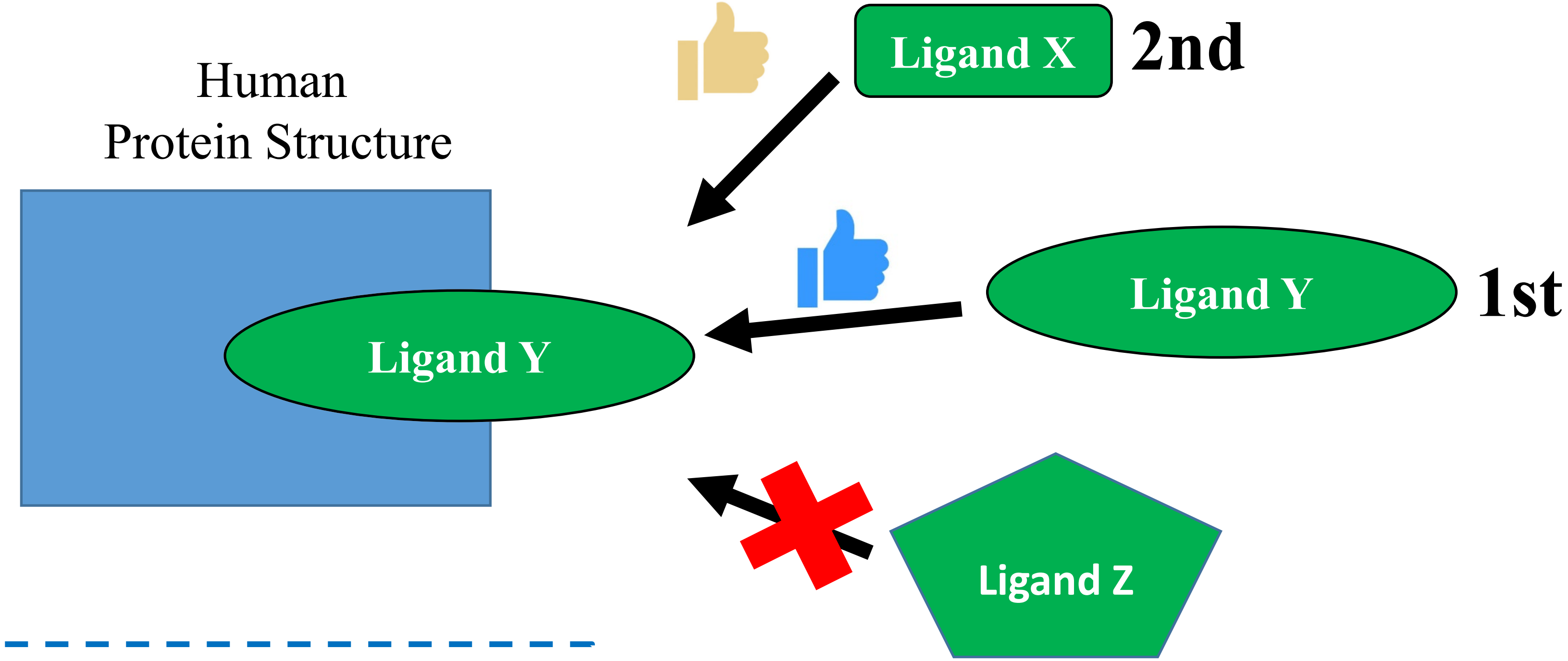
Factors that make a species sensitive

- Exposure
- Dose
- ADME
- Target receptor availability
- Life stage
- Life history
- etc.
- etc.

Advances in Drug Discovery/Development

(COVID-19 has led to advances)

Structure derived
from X-ray
crystallography



Human
Protein Structure

Ligand Y

Ligand X

2nd

Ligand Y

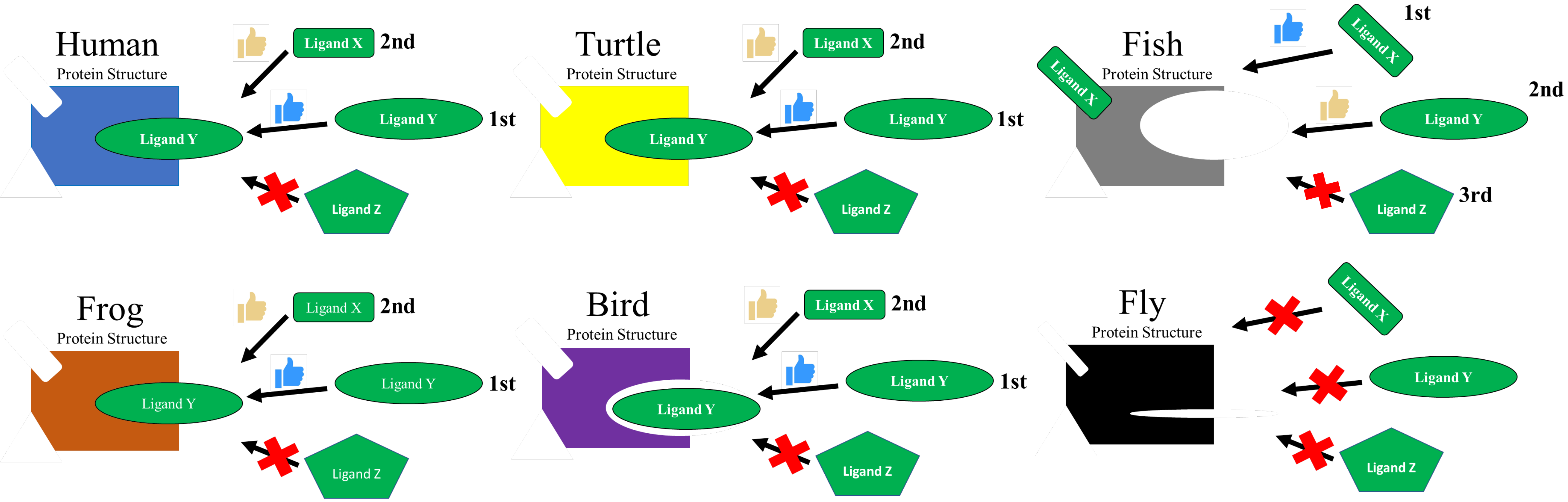
1st

Ligand Z

<u>Bioinformatics Toolbox:</u>

Molecular modeling
Molecular docking
Virtual screening
Molecular dynamic simulations

Application to Species Extrapolation



Bioinformatics Toolbox:

Molecular modeling

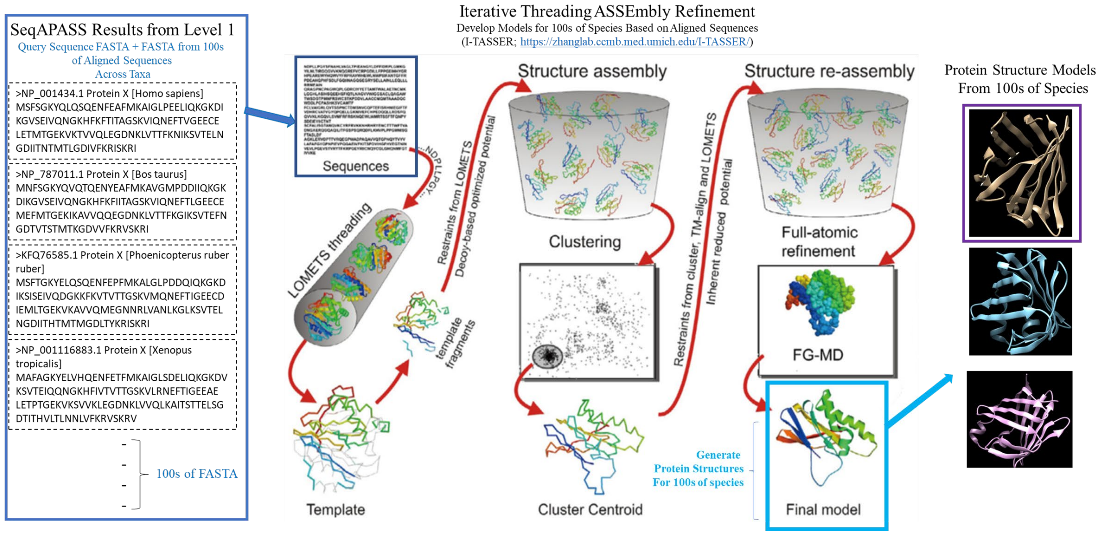
Molecular docking

Virtual screening

Molecular dynamic simulations

Thousands/Millions/Billions
of
Chemicals

Expanding the SeqAPASS Pipeline: I-TASSER and TM-Align



SeqAPASS Results from Level 1

Query Sequence FASTA + FASTA from 100s of Aligned Sequences Across Taxa

```
>NP_001434.1 Protein X [Homo sapiens]
MSFSGKYQLQSQENFEAFMKAIGLPEELIQKGKDI
KGVSEIVQNGKHFKFTITAGSKVIQNEFTVGEECE
LETMTGEKVKTVVQLEGDNKLVTTFKNIKSVTELN
GDIITNTMTLGDIVFKRISKRI
```

```
>NP_787011.1 Protein X [Bos taurus]
MNFSGKYQVQTQENYEAFMKAVGMPDDIIQKGK
DIKGVSEIVQNGKHFKFIITAGSKVIQNEFTLGEECE
MEFMTGEKIKAVVQQEGDNKLVTTFKGIKSVTEFN
GDTVTSTMTKGDVVFKRVSKRI
```

```
>KFQ76585.1 Protein X [Phoenicopterus ruber
ruber]
MSFTGKYELQSQENFEPFMKALGLPDDQIQKGKD
IKSISEIVQDGKKFKVTVTTGSKVMQNEFTIGEECD
IEMLTGEKVKAVVQMEGNNRLVANLKGLKSVTEL
NGDIITHTMTMGDLTYKRISKRI
```

```
>NP_001116883.1 Protein X [Xenopus
tropicalis]
MAFAGKYELVHQENFETFMKAIGLSDELIQKGKDV
KSVTEIQQNGKHFIVTVTTGSKVLRNEFTIGEEAE
LETPTGEKVKSVVKLEGDNKLVVQLKAITSTTELSG
DTITHVLTLNNLVFKRVSKRV
```

100s of FASTA

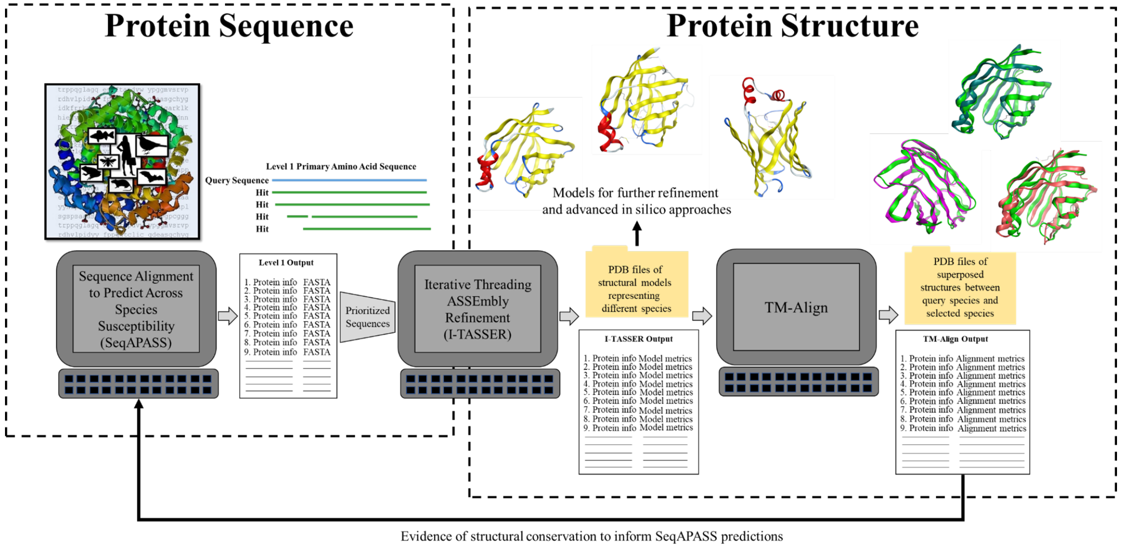
Iterative Threading ASSEmbly Refinement

Develop Models for 100s of Species Based on Aligned Sequences

(I-TASSER; https://zhanglab.ccmb.med.umich.edu/I-TASSER/)

Generate Protein Structures For 100s of species

Protein Structure Models From 100s of Species



Protein Sequence

Protein Structure

Sequence Alignment to Predict Across Species Susceptibility (SeqAPASS)

Level 1 Output

1. Protein info FASTA
2. Protein info FASTA
3. Protein info FASTA
4. Protein info FASTA
5. Protein info FASTA
6. Protein info FASTA
7. Protein info FASTA
8. Protein info FASTA
9. Protein info FASTA

Prioritized Sequences

Iterative Threading ASSEmbly Refinement (I-TASSER)

PDB files of structural models representing different species

Models for further refinement and advanced in silico approaches

I-TASSER Output

1. Protein info Model metrics
2. Protein info Model metrics
3. Protein info Model metrics
4. Protein info Model metrics
5. Protein info Model metrics
6. Protein info Model metrics
7. Protein info Model metrics
8. Protein info Model metrics
9. Protein info Model metrics

TM-Align

PDB files of superposed structures between query species and selected species

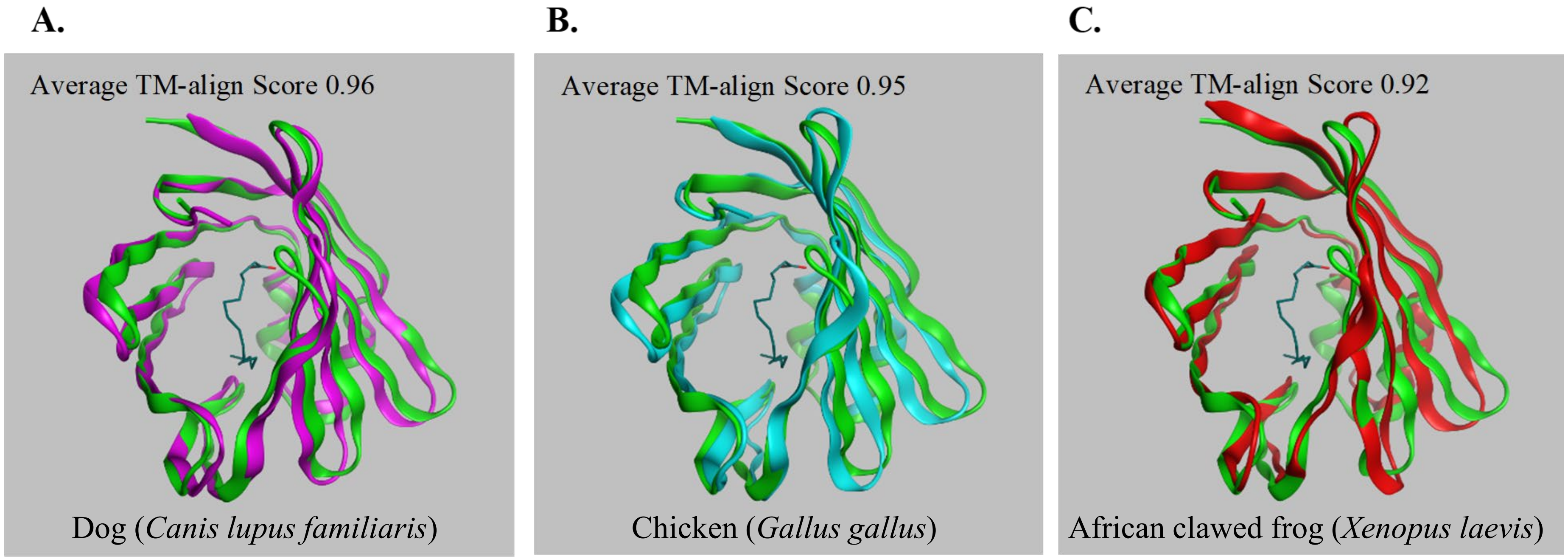
TM-Align Output

1. Protein info Alignment metrics
2. Protein info Alignment metrics
3. Protein info Alignment metrics
4. Protein info Alignment metrics
5. Protein info Alignment metrics
6. Protein info Alignment metrics
7. Protein info Alignment metrics
8. Protein info Alignment metrics
9. Protein info Alignment metrics

Evidence of structural conservation to inform SeqAPASS predictions

Query Species:

Human
Androgen receptor
Ligand Binding Domain



Environmental
Toxicology and Chemistry

Environmental Toxicology | Full Access

From Protein Sequence to Structure: The Next Frontier in Cross-Species Extrapolation for Chemical Safety Evaluations

Carlie A. LaLone, Donovan J. Blatz, Marissa A. Jensen, Sara M. F. Vliet, Sally Mayasich, Kali Z. Mattingly, Thomas R. Transue, Wilson Melendez, Audrey Wilkinson, Cody W. Simmons, Carla Ng ... See all authors

First published: 16 December 2022 | https://doi.org/10.1002/etc.5537

United States
Environmental Protection
Agency

Acknowledgements

U.S. EPA, ORD
- Marissa Brickley (University of Minnesota Duluth)
- Ryan Staub (ORISE)
- Sally Mayasich (past University of Wisconsin 2023)
- Maxwell Botz (past ORISE2 2023)
- Peter Schumann (past ORISE 2023)
- Monique Hazemi (past ORISE 2023)
- Sara Vliet (past ORISE 2021)
- Donovan Blatz (past ORISE 2021)
- Jon Doering (past NRC 2018)
- Colin Finnegan (past ORISE 2018)

GDIT
- Cody Simmons
- Audrey Wilkinson
- Wilson Menendez
- Thomas Transue (past GDIT 2022)

SeqAPASS v7.1

Choose Search Type: By Species

LaLone.Carlie@epa.gov

https://seqapass.epa.gov/seqapass/