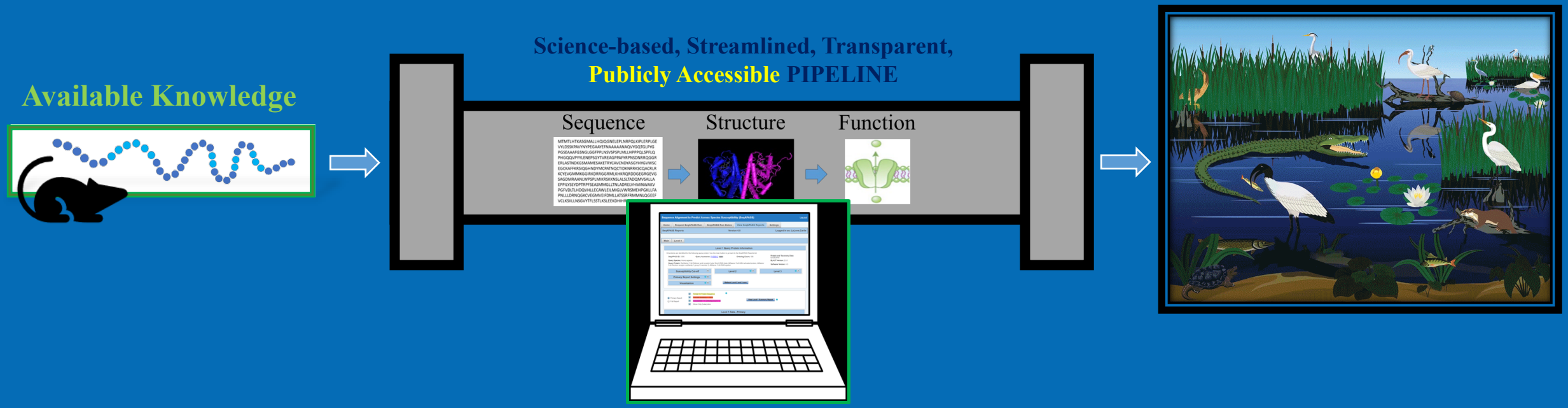
EPA
United States
Environmental Protection
Agency

Advancing the SeqAPASS Pipeline from Sequence to Structure to Evaluate Protein Conservation for Species Extrapolation

Carlie A. LaLone



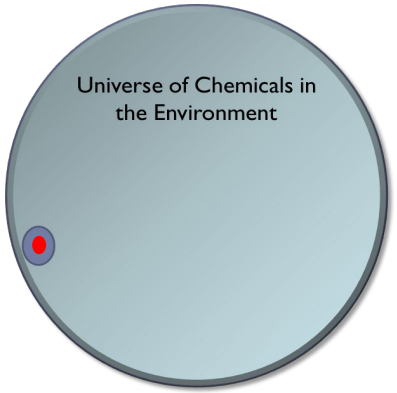
Office of Research and Development
Center for Computational Toxicology and Exposure, Great Lakes Toxicology and Ecology Division

The views expressed in this presentation are those of the authors and do not necessarily reflect the views or policies of the US EPA

Chemical Safety Evaluation

- Protect human health and the environment
  - Ensure that chemicals in the marketplace are reviewed for safety

- Challenging mission:
  - Tens of thousand of chemicals are currently in use and hundreds are introduced annually
  - Many have not been thoroughly evaluated for potential risk to human health and the environment
    - *Chemicals tested across species: Even more sparse*



EPA United States Environmental Protection Agency

Reduce Animal Testing at the US EPA

- EPA Administrator Andrew Wheeler signed directive (Sept. 10th 2019) to reduce animal testing
  - Calls for the Agency to:
    - Reduce its request for, and funding of, mammal studies by 30% by 2025
      - That is ~5 years from today!
    - Eliminate all mammal study requests and funding by 2035
      - That is ~15 years from today

How do we get there?

NAMs

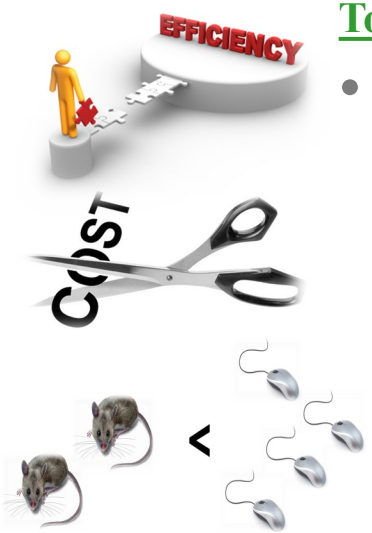
Transformation of Toxicity Testing

Historically:

Whole animal test

- **Observe Toxic Outcome**
  - **Examples**
    - **tumor development**
    - **mortality**

  Resource intensive

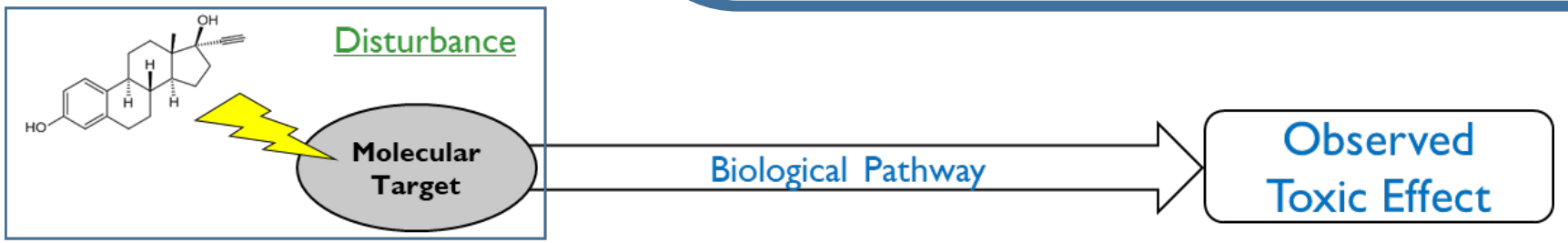


Toxicity Testing in the 21st Century:

- ***In vitro* and *in silico* methods**
  - **Pathway-based approaches**
  - **Focus on disturbance of the biological pathway**
    - **Predictive of the observable toxic effects**

New Approach Methods (NAMs)

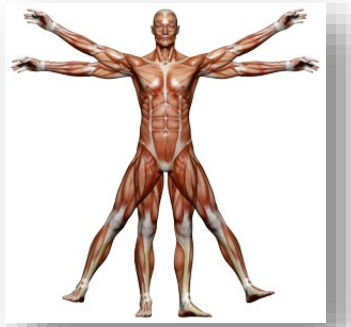
- Informatics
- High throughput
- Systems biology
- OMICs



Enabled by evolution of the science and technology

Model Organisms for Toxicity Testing

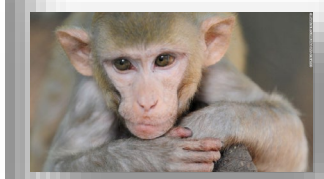
- **Assumed** that sensitivity of species to a chemical is a function of their relatedness



Cannot Test

≅

- **Human Health Risk Assessment**

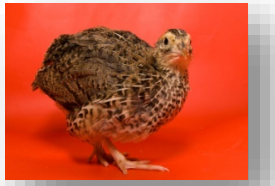


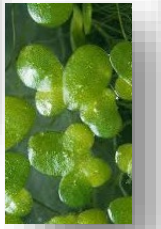
- **Ecological Risk Assessment**

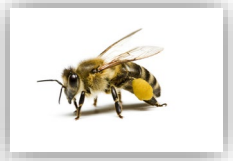
Use of Surrogates

Cannot Test

≅







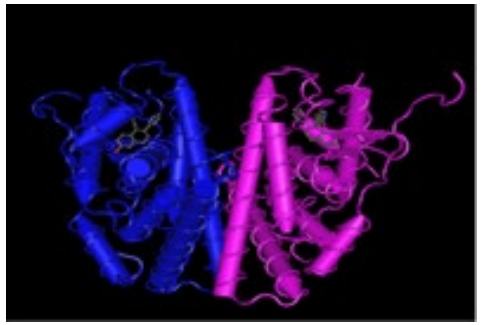
Representative species across a diversity of organism classes

Sequence

MTMTLHTKASGMALLHQIQGNELEPLNRPQLKIPLERPLGE
VYLDSSKPAVYNYPEGAAYEFNAAAAANAQVYGQTGLPYG
PGSEAAAFGSNGLGGFPPLNSVSPSPLMLLHPPPQLSPFLQ
PHGQQVPYYLENEPSGYTVREAGPPAFYRPNSDNRRQGGR
ERLASTNDKGSMAMESAKETRYCAVCNDYASGYHYGVWSC
EGCKAFFKRSIQGHNDYMCPATNQCTIDKNRRKSCQACRLR
KCYEVGMMKGGIRKDRRGGRMLKHKRQRDDGEGRGEVG
SAGDMRAANLWPSPLMIKRSKKNSLALSLTADQMVSALLA
EPPILYSEYDPTRPFSEASMMGLLTNLADRELVHMINWAKV
PGFVDLTLHDQVHLLECAWLEILMIGLVWRSMEHPGKLLFA
PNLLLDRNQGKCVEGMVEIFDMLLATSSRFRMMNLQGEEF
VCLKSIILLNSGVYTFLSSTLKSLEEKDHIHRVLDKITDTLIHLM

Yes or No
Susceptible or Not Susceptible

Structure



Structural-based
comparisons of similarity
Predicted binding affinity

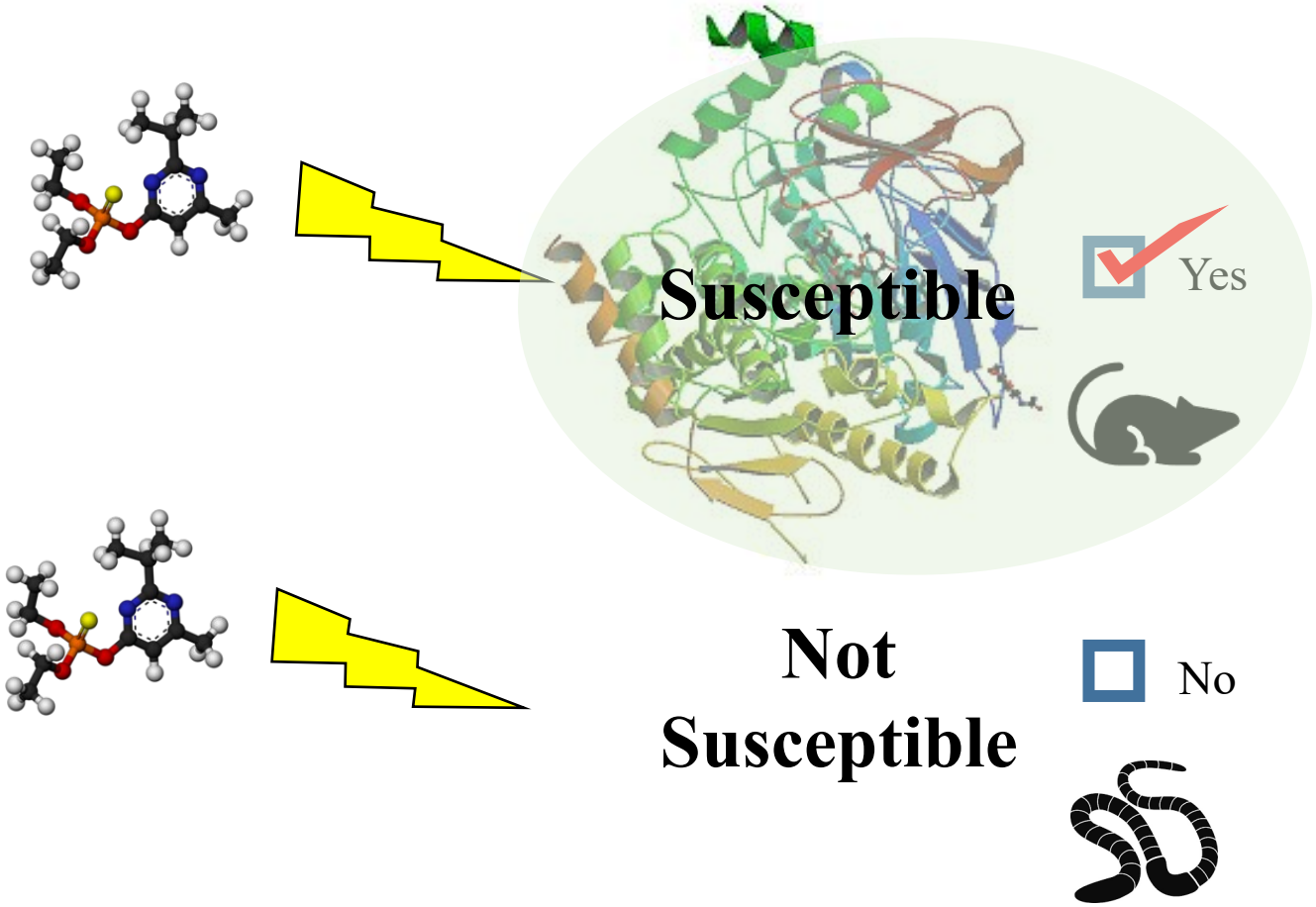
Function

Improvements in bioinformatics

Considering chemical sensitivity?

Factors that make a species sensitive

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



Simple question to address:

Is the known chemical target available in a species for a chemical to act upon?

Yes or No

Likely susceptible or Not likely susceptible (at least through the known mechanism)

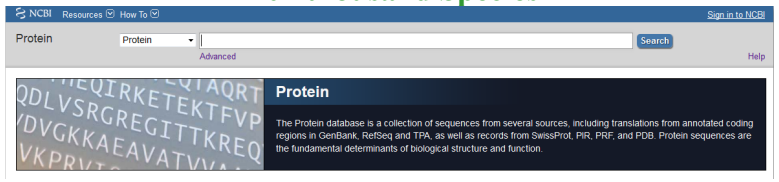
United States Environmental Protection Agency

New Approach Methods: Species Extrapolation

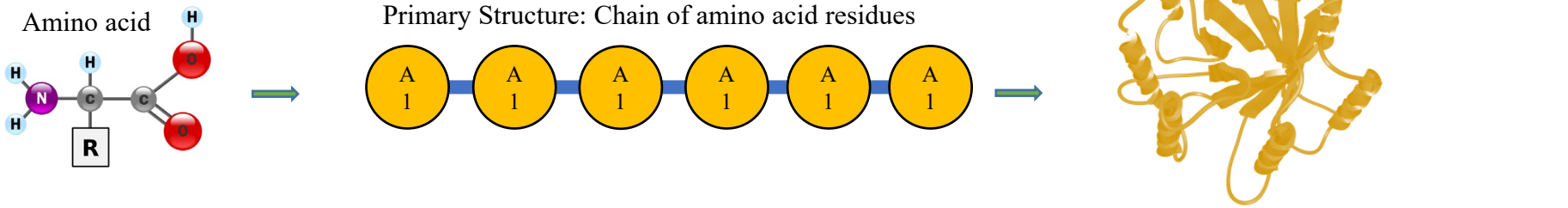
New tools and technologies have emerged

- Improved sequencing technologies
- Large databases of sequence data

As of this week
~172 million Proteins
~101 thousand Species



- **Focus on the molecular machine: The Protein**
  - **Large biomolecule assembled from amino acids encoded in genes**



  - **Many functions (e.g., catalyze reactions, structural/mechanical functions, cell signaling, immune response, etc.)**

- **Evaluate protein similarity between species**
  - **Moving away from empirical testing and qualitative understanding of molecular target (protein) conservation to quantitative measures**

EPA United States Environmental Protection Agency

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

OXFORD SOT Society of Toxicology www.toxsci.oxfordjournals.org

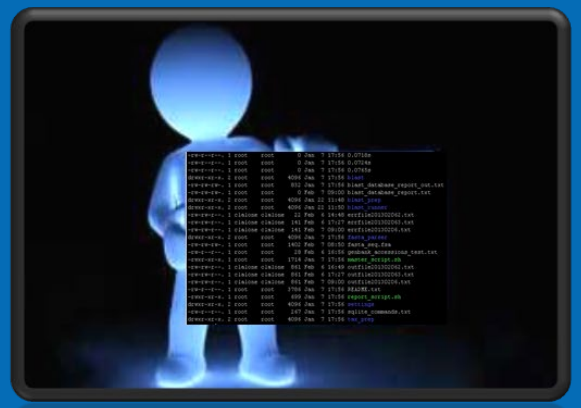
doi: 10.1093/toxsci/kfw119
Advance Access Publication Date: June 30, 2016
Research article

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

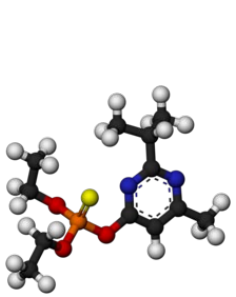
Carlie A. LaLone,*,[1] Daniel L. Villeneuve,* David Lyons,† Henry W. Helgen,‡ Serina L. Robinson,§,[2] Joseph A. Swintek,¶ Travis W. Saari,* and Gerald T. Ankley*

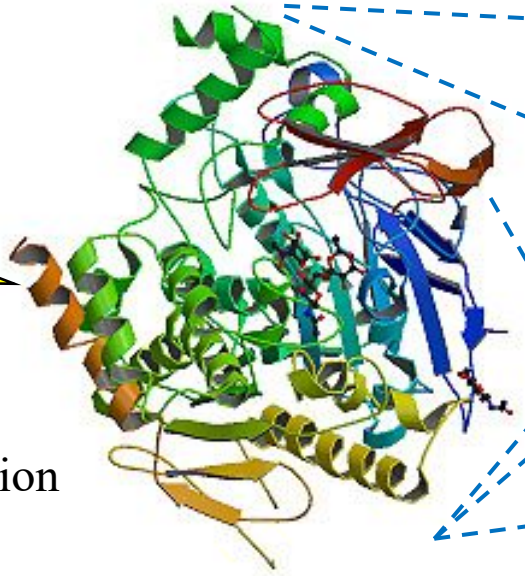
https://seqapass.epa.gov/seqapass/

Sequence Alignment to Predict Across Species Susceptibility (SeqAPASS)



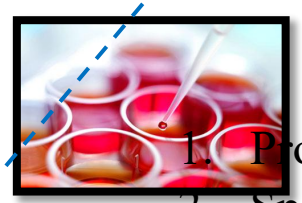
What information is required for a SeqAPASS query?





Chemical-Protein Interaction

Knowledge of a sensitive or targeted species

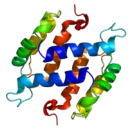


Knowledge of the model organism used in an *in vitro* assay

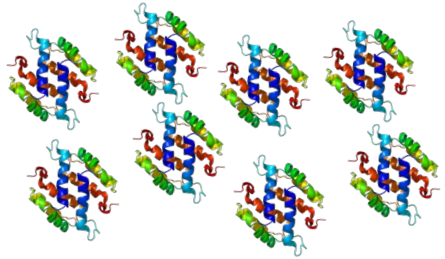
1. Protein
2. Species

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

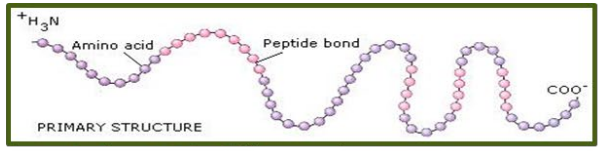
Greater similarity = Greater likelihood that chemical can act on the protein

Line of Evidence: Predict Potential Chemical Susceptibility Across Species

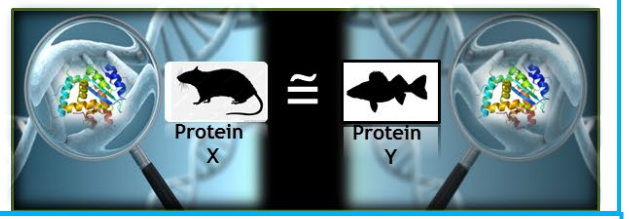
Strategic Automated Approach for Assessing Protein (Molecular Target) Similarity

Level 1

Primary Amino Acid Sequence Alignments

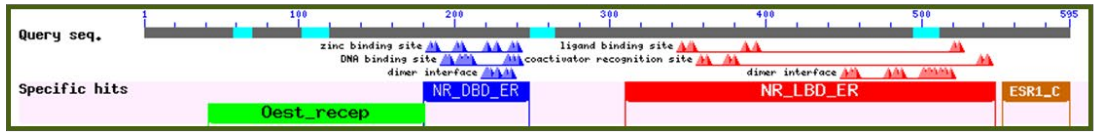


Ortholog Candidate Identification (RBH)



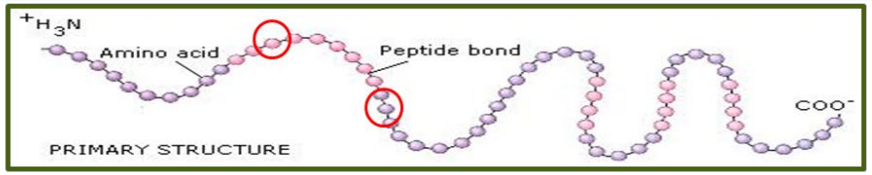
Level 2

Conserved Functional Domain Alignments



Level 3

Individual Amino Acid Residue Queries



Tertiary Protein Structure Considerations

Low Level of Complexity

High level of Complexity

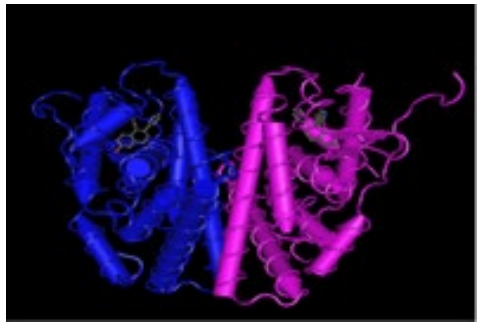
Flexibility to use Existing Knowledge

Sequence

MTMTLHTKASGMALLHQIQGNELEPLNRPQLKIPLERPLGE
VYLDSSKPAVYNYPEGAAYEFNAAAAANAQVYGQTGLPYG
PGSEAAAFGSNGLGGFPPLNSVSPSPLMLLHPPPQLSPFLQ
PHGQQVPYYLENEPSGYTVREAGPPAFYRPNSDNRRQGGR
ERLASTNDKGSMAMESAKETRYCAVCNDYASGYHYGVWSC
EGCKAFFKRSIQGHNDYMCPATNQCTIDKNRRKSCQACRLR
KCYEVGMMKGGIRKDRRGGRMLKHKRQRDDGEGRGEVG
SAGDMRAANLWPSPLMIKRSKKNSLALSLTADQMVSALLA
EPPILYSEYDPTRPFSEASMMGLLTNLADRELVHMINWAKV
PGFVDLTLHDQVHLLECAWLEILMIGLVWRSMEHPGKLLFA
PNLLLDRNQGKCVEGMVEIFDMLLATSSRFRMMNLQGEEF
VCLKSIILLNSGVYTFLSSTLKSLEEKDHIHRVLDKITDTLIHLM

Yes or No
Susceptible or Not Susceptible

Structure



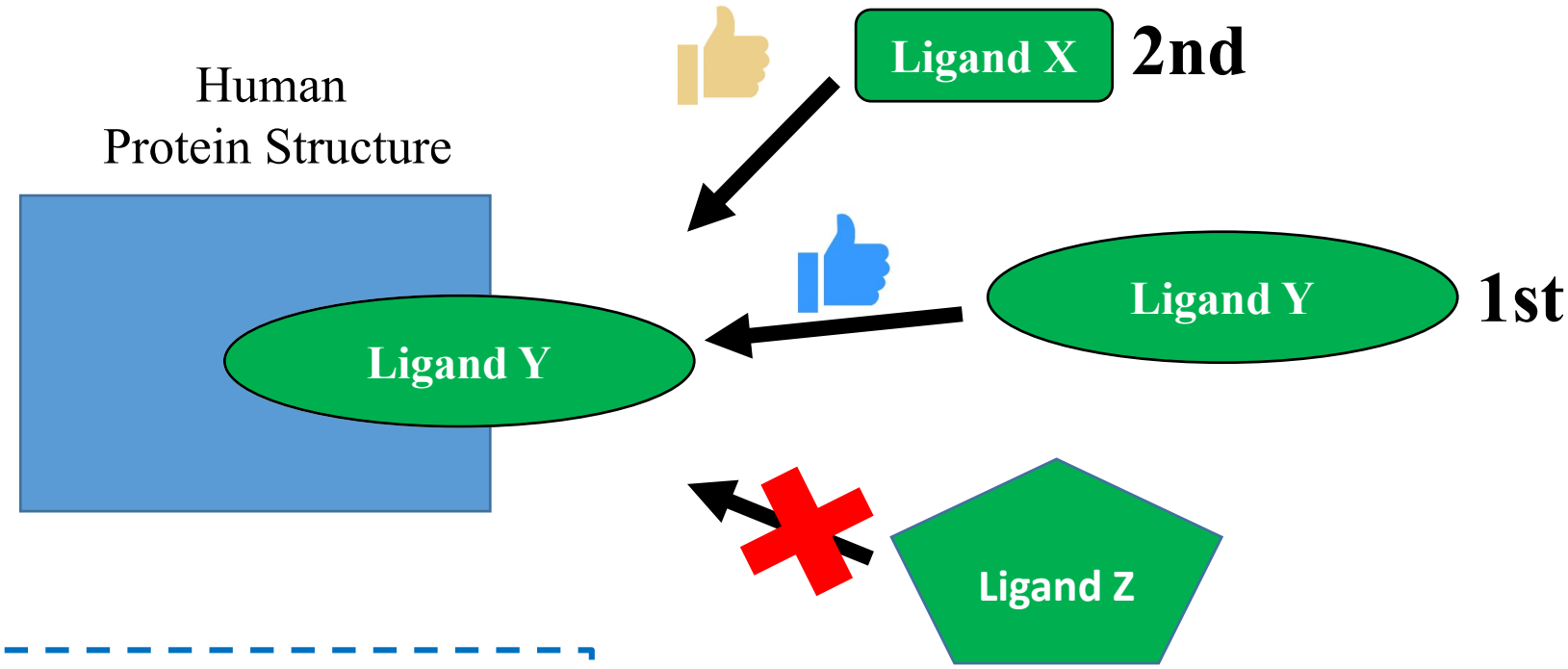
Structural-based
comparisons of similarity
Predicted binding affinity

Function

Improvements in bioinformatics

Advances in Drug Discovery/Development

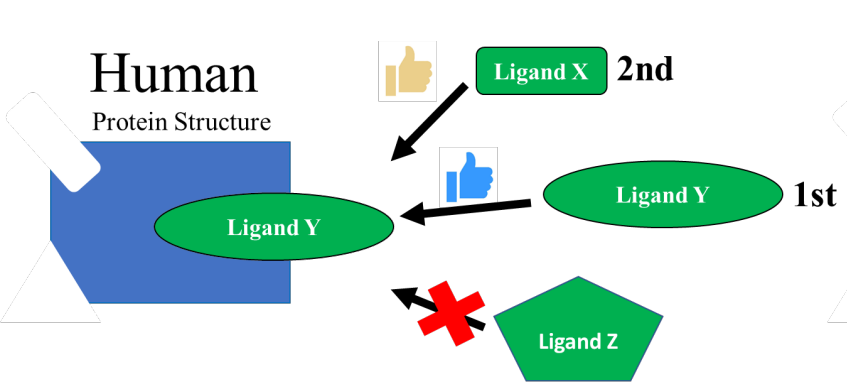
Structure derived from X-ray crystallography

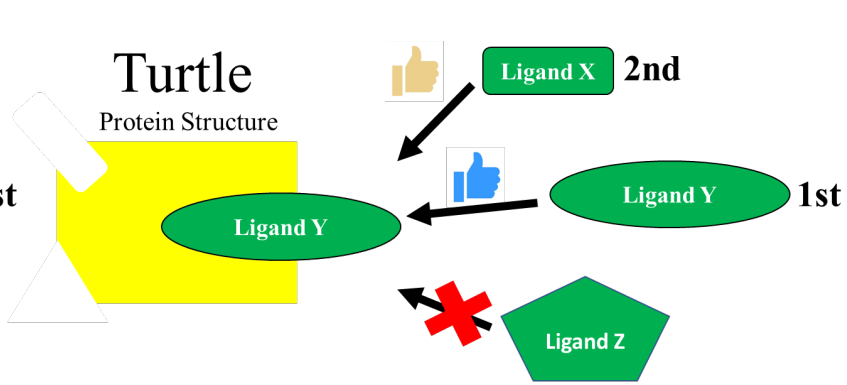


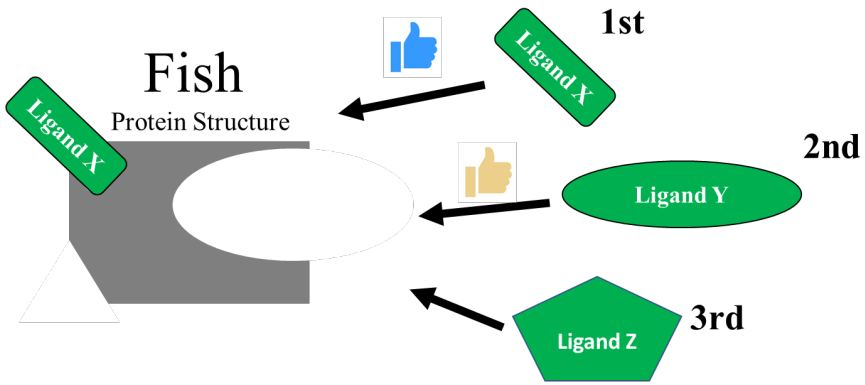
Bioinformatics Toolbox:

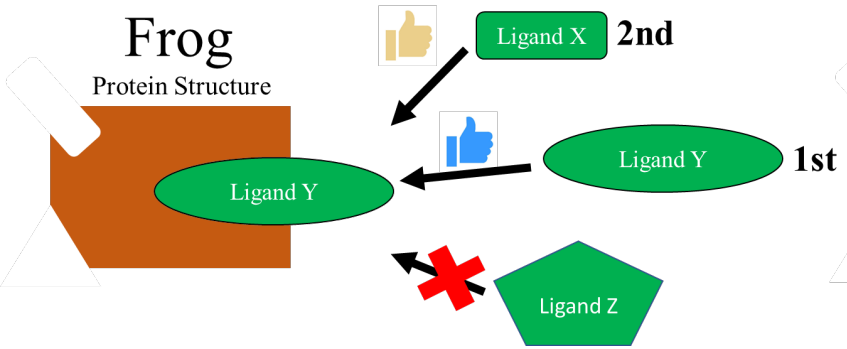
Molecular modeling
Molecular docking
Virtual screening
Molecular dynamic simulations

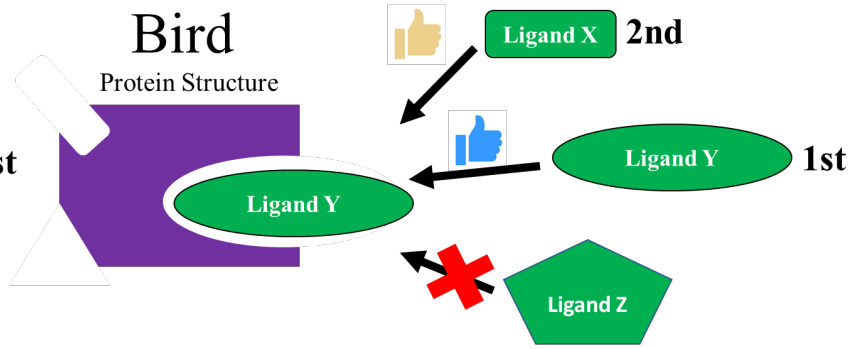
Application to Species Extrapolation

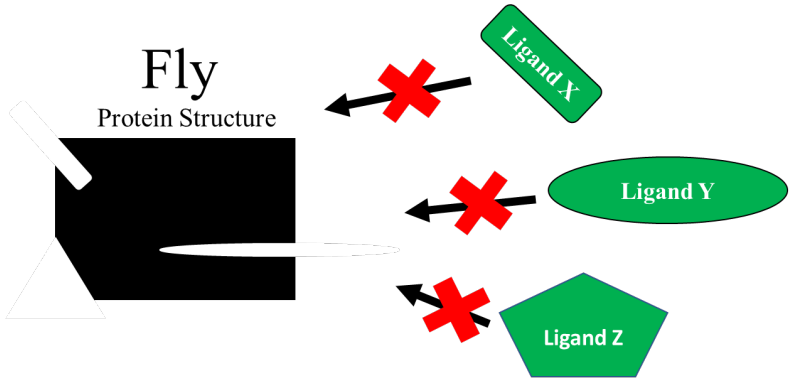










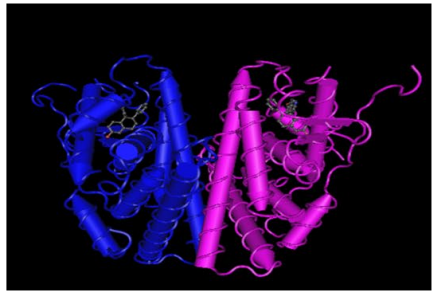


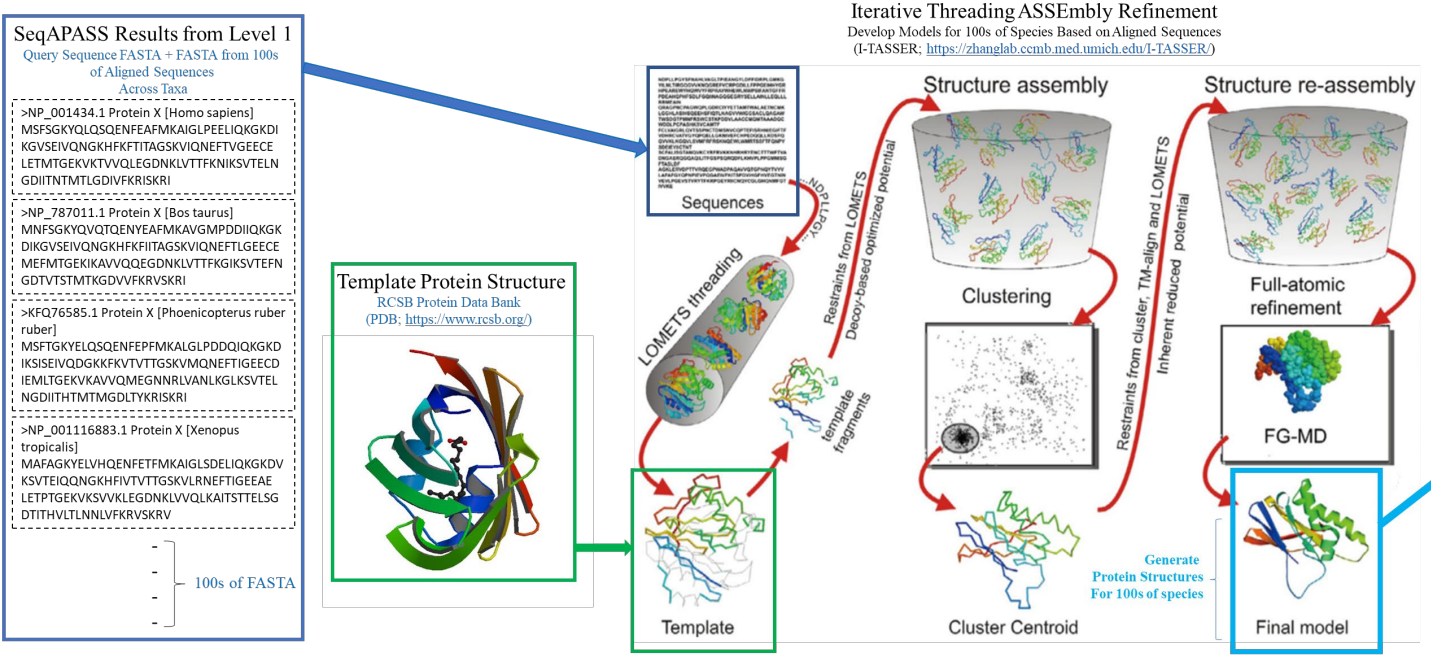
Bioinformatics Toolbox:

Molecular modeling
Molecular docking
Virtual screening
Molecular dynamic simulations

Sequence

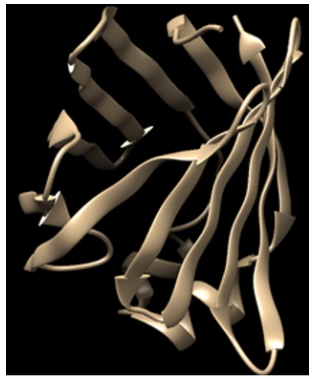
MTMTLHTKASGMALLHQIQGNELEPLNRPQLKIPLERPLGE
VYLDSSKPAVYNYPEGAAYEFNAAAAANAQVYGQTGLPYG
PGSEAAAFGSNGLGGFPPLNSVSPSPLMLLHPPPQLSPFLQ
PHGQQVPYYLENEPSGYTVREAGPPAFYRPNSDNRRQGGR
ERLASTNDKGSMAMESAKETRYCAVCNDYASGYHYGVWSC
EGCKAFFKRSIQGHNDYMCPATNQCTIDKNRRKSCQACRLR
KCYEVGMMKGGIRKDRRGGRMLKHKRQRDDGEGRGEVG
SAGDMRAANLWPSPLMIKRSKKNSLALSLTADQMVSALLA
EPPILYSEYDPTRPFSEASMMGLLTNLADRELVHMINWAKV
PGFVDLTLHDQVHLLECAWLEILMIGLVWRSMEHPGKLLFA
PNLLLDRNQGKCVEGMVEIFDMLLATSSRFRMMNLQGEEF
VCLKSIILLNSGVYTFLSSTLKSLEEKDHIHRVLDKITDTLIHLM





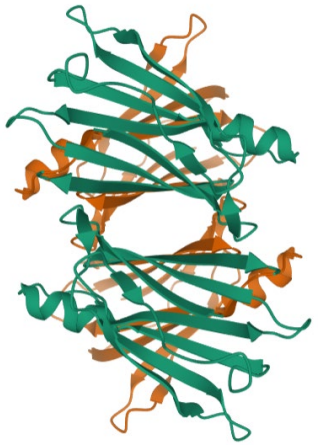
Graphic Modified from Zhang et al., 2019 I-TASSER gateway: A protein structure and function prediction server powered by XSEDE Figure 1

Species	C-Score	TM-Score	RMSD
Human	1.68	0.95+/-0.05	1.3+/-1.3
Cattle	1.64	0.94+/-0.05	1.4+/-1.3
Flamingo	1.60	0.94+/-0.05	1.5+/-1.3
etc.	-	-	-
etc.	-	-	-
etc.	-	-	-



Liver Fatty Acid Protein

Comparison to Human LFABP	Chain 1 (TM-score)	Chain 2 (TM-score)	Aligned Length	RMSD	n_identical/ n_aligned	SeqAPASS % similarity
Human	1.00000	1.00000	127	0.00	1.000	100
Orangutan	0.99766	0.885385	127	0.20	0.976	97.71
Rat	0.99696	0.99696	127	0.23	0.827	85.04
Cattle	0.99761	0.99761	127	0.21	0.811	85.04
Water Flea	0.90965	0.89030	126	1.34	0.397	27.32
Marine worm	0.92609	0.89960	126	1.18	0.349	27.02
Round worm	0.89736	0.47865	126	1.45	0.286	18.32
Fruit Fly	0.95687	0.93550	126	0.90	0.278	18.93



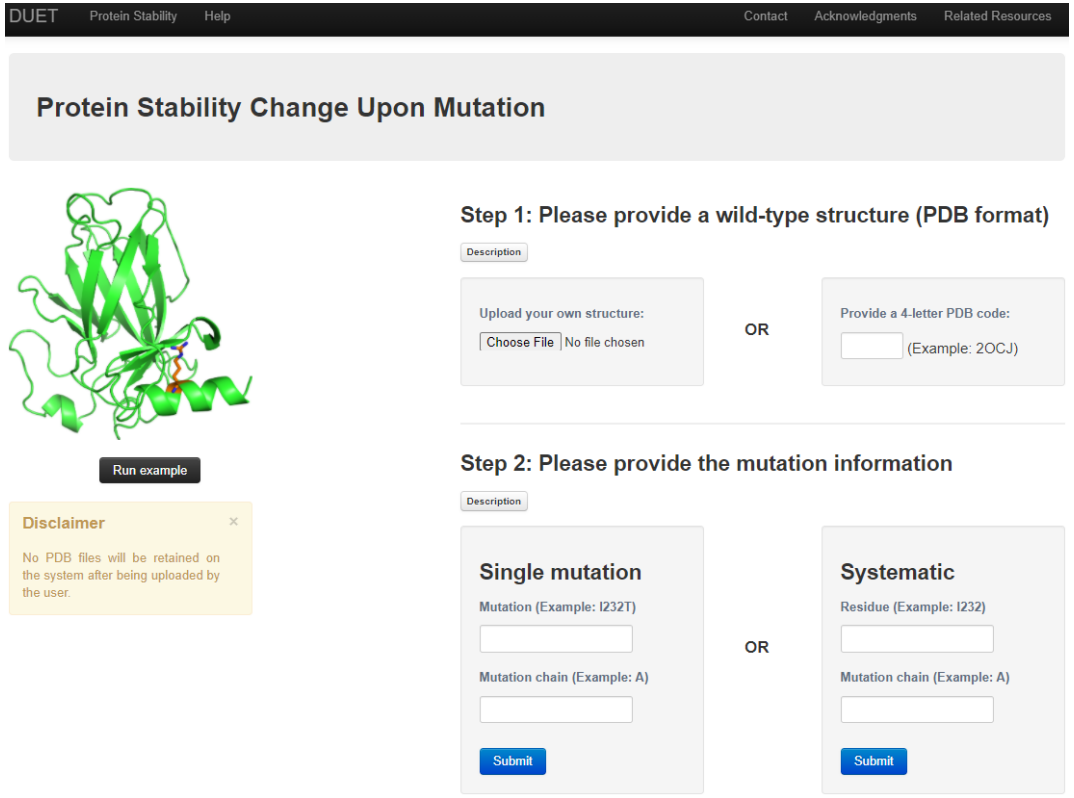
Transthyretin (TTR)

Comparison to Human TTR	Chain 1 (TM-Score)	Chain 2 (TM-Score)	Aligned Length	RMSD	N_identical/n _aligned	SeqAPASS % similarity
Human	1.00000	1.00000	147	0.00	1.000	100
Orangutan	0.86434	0.86434	144	1.86	0.861	89.81
Cattle	0.86403	0.86403	144	2.02	0.743	86.11
Red Deer	0.87384	0.87384	144	1.94	0.729	83.57
Bar Tailed Godwit	0.87155	0.85562	145	2.08	0.662	74.65
Three Toed Box Turtle	0.85385	0.83860	144	1.99	0.650	66.88
Zebrafish	0.84586	0.83580	145	2.18	0.497	48.53
Acorn worm	0.79295	0.80791	132	2.03	0.333	28.79

Capturing structural similarity metrics to understand conservation of chemical targets across species

United States Environmental Protection Agency

DUET | Instituto René Rachou FIOCRUZ MINAS | The University of Melbourne | University of Cambridge



Combine SeqAPASS predictions to structure
Level 3 of SeqAPASS – identify amino acid differences across species
DUET predict stability changes from amino acid differences across species

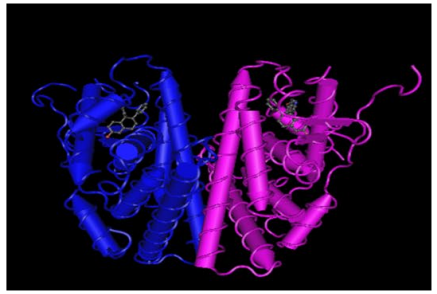
Human Amino Acid Position	Type 1 Primates, Ruminants, Whales/dolphins	Type 2 Rodents and other mammals, Fish, Amphibians, Testudines	Type 3 Aves, Lepidosauria Chondrichthyes	Type 4 Crocodylia	SeqAPASS Level 3 Prediction of Similar to Human LFABP Template	Mutation in DUET	Stability Change from DUET (ΔΔG, kcal/mol)
50	Phenylalanine (F)			Phenylalanine	Yes		
		Valine (V)	Valine (V)		No	F50V	-1.196 (Destabilizing)
		Isoleucine (I)	Isoleucine (I)		No	F50I	-0.808 (Destabilizing)
		Leucine (L)	Leucine (L)		No	F50L	-0.893 (Destabilizing)
54	Alanine (A)				Yes		
		Threonine (T)	Threonine	Threonine	No	A54T	-0.195 (Destabilizing)
81	Threonine (T)			Threonine	Yes		
		Alanine (A)	Alanine		No	T81A	-0.749 (Destabilizing)
		Glycine (G)			No	T81G	-0.023 (Destabilizing)
93	Threonine (T)	Threonine			Yes		
		Valine			Yes	T93V	0.031 (Stabilizing)
			Alanine		No	T93A	-1.004 (Destabilizing)
97	Asparagine (N)				Yes		
		Glycine	Glycine	Glycine	No	N97G	0.521 (Stabilizing)

Combined sequence and structure: another line of evidence toward conservation

Sequence

MTMTLHTKASGMALLHQIQGNELEPLNRPQLKIPLERPLGE
VYLDSSKPAVYNYPEGAAYEFNAAAAANAQVYGQTGLPYG
PGSEAAAFGSNGLGGFPPLNSVSPSPLMLLHPPPQLSPFLQ
PHGQQVPYYLENEPSGYTVREAGPPAFYRPNSDNRRQGGR
ERLASTNDKGSMAMESAKETRYCAVCNDYASGYHYGVWSC
EGCKAFFKRSIQGHNDYMCPATNQCTIDKNRRKSCQACRLR
KCYEVGMMKGGIRKDRRGGRMLKHKRQRDDGEGRGEVG
SAGDMRAANLWPSPLMIKRSKKNSLALSLTADQMVSALLA
EPPILYSEYDPTRPFSEASMMGLLTNLADRELVHMINWAKV
PGFVDLTLHDQVHLLECAWLEILMIGLVWRSMEHPGKLLFA
PNLLLDRNQGKCVEGMVEIFDMLLATSSRFRMMNLQGEEF
VCLKSIILLNSGVYTFLSSTLKSLEEKDHIHRVLDKITDTLIHLM

Structure



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

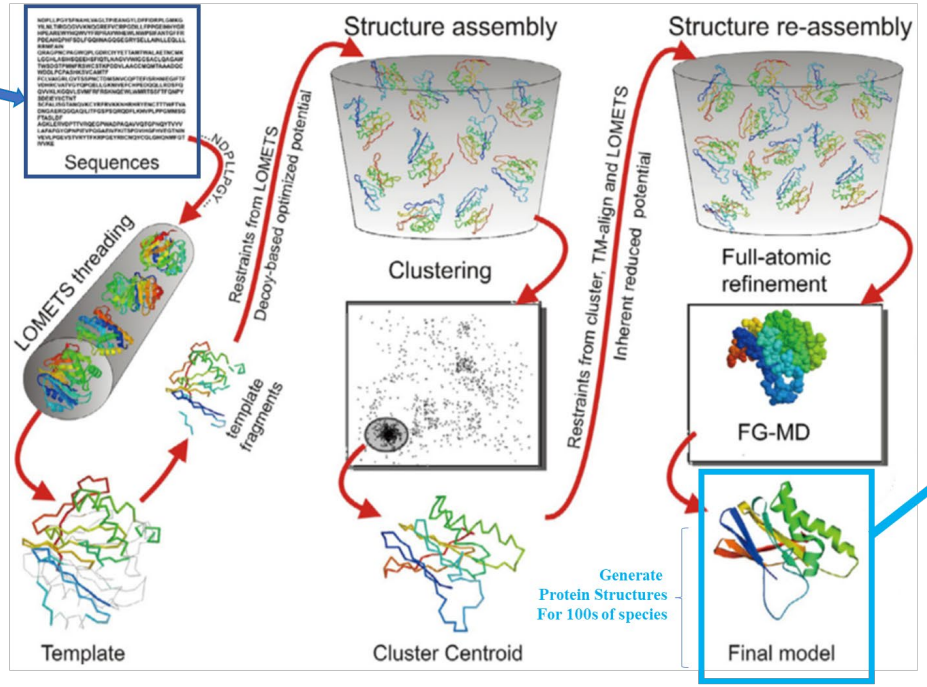
>KFQ76585.1 Protein X [Phoenicopterus ruber ruber]
MSFTGKYELQSQENFEPFMKALGLPDDQIQKGKD
IKSISEIVQDGKKFKVTVTTGSKVMQNEFTIGEECD
IEMLTGEKVKAVVQMEGNNRLVANLKGLKSVTEL
NGDIITHTMTMGDLTYKRISKRI

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

100s of FASTA

Iterative Threading ASSEmbly Refinement

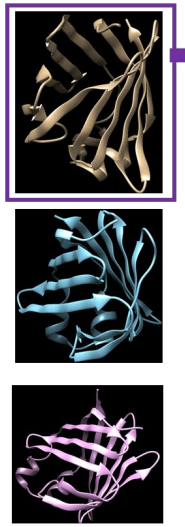
Develop Models for 100s of Species Based on Aligned Sequences
(I-TASSER; https://zhanglab.ccmb.med.umich.edu/I-TASSER/)



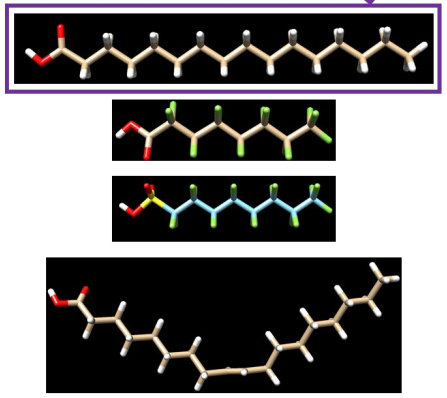
UCSF Chimera

DockPrep Structures and Minimize Ligands

Protein Structure Models From 100s of Species



Ligands of Interest for Docking



AutoDock Vina

Dock Multiple Ligands to Protein Structures

Collect Predicted Binding Affinity

S	Score	RMSD lb.	RMSD u.b.	HBonds (all)	HBond Ligand Atoms	HBond Receptor Atoms
V	-7.1	0.0	0.0	0	0	0
V	-7.0	1.212	2.436	0	0	0
V	-7.0	2.148	6.837	1	1	1
V	-6.9	1.128	2.04	0	0	0
V	-6.9	4.472	7.133	0	0	0
V	-6.7	3.27	7.552	0	0	0
V	-6.7	2.637	3.461	2	2	2
V	-6.6	1.572	3.516	0	0	0
V	-6.6	1.725	3.368	0	0	0

Chimera Model #3.1

```
REMARK VINA RESULT:      -7.1      0.000      0.000
REMARK  15 active torsions:
REMARK  status: ('A' for Active; 'I' for Inactive)
REMARK    1  A    between atoms: C2_2  and  C3_3
REMARK    2  A    between atoms: C3_3  and  C4_4
REMARK    3  A    between atoms: C4_4  and  C5_5
REMARK    4  A    between atoms: C5_5  and  C6_6
REMARK    5  A    between atoms: C6_6  and  C7_7
REMARK    6  A    between atoms: C7_7  and  C8_8
REMARK    7  A    between atoms: C8_8  and  C9_9
REMARK    8  A    between atoms: C10_10  and  C9_9
REMARK    9  A    between atoms: C10_10  and  C11_11
REMARK   10  A    between atoms: C11_11  and  C12_12
REMARK   11  A    between atoms: C12_12  and  C13_13
REMARK   12  A    between atoms: C13_13  and  C14_14
REMARK   13  A    between atoms: C14_14  and  C15_15
REMARK   14  A    between atoms: C15_15  and  C16_16
REMARK   15  A    between atoms: C16_16  and  O2_18
```

Graphic Modified from Zhang et al., 2019 I-TASSER gateway: A protein structure and function prediction server powered by XSEDE Figure 1

Predicting Binding Affinity

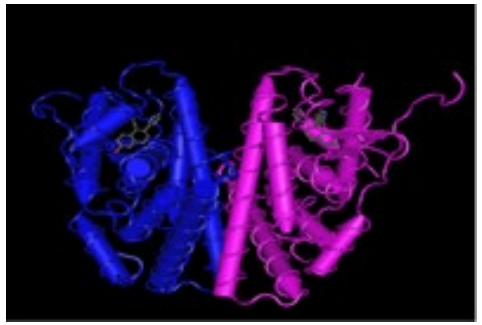

Application of SeqAPASS
MIE
KE
KE
KE
AO

Sequence

MTMTLHTKASGMALLHQIQGNELEPLNRPQLKIPLERPLGE
VYLDSSKPAVYNYPEGAAYEFNAAAAANAQVYGQTGLPYG
PGSEAAAFGSNGLGGFPPLNSVSPSPLMLLHPPPQLSPFLQ
PHGQQVPYYLENEPSGYTVREAGPPAFYRPNSDNRRQGGR
ERLASTNDKGSMAMESAKETRYCAVCNDYASGYHYGVWSC
EGCKAFFKRSIQGHNDYMCPATNQCTIDKNRRKSCQACRLR
KCYEVGMMKGGIRKDRRGGRMLKHKRQRDDGEGRGEVG
SAGDMRAANLWPSPLMIKRSKKNSLALSLTADQMVSALLA
EPPILYSEYDPTRPFSEASMMGLLTNLADRELVHMINWAKV
PGFVDLTLHDQVHLLECAWLEILMIGLVWRSMEHPGKLLFA
PNLLLDRNQGKCVEGMVEIFDMLLATSSRFRMMNLQGEEF
VCLKSIILLNSGVYTFLSSTLKSLEEKDHIHRVLDKITDTLIHLM

Yes or No
Susceptible or Not Susceptible

Structure



Structural-based
comparisons of similarity
Predicted binding affinity

Function

Improvements in bioinformatics

Acknowledgements

U.S. EPA, ORD

Marissa Jensen (University of Minnesota Duluth)
Sally Mayasich (ORISE)
Sara Vliet (ORISE)
Donovan Blatz (ORISE)
Jon Doering (U of Lethbrigde)
Colin Finnegan (Iowa State University)))

GDIT

Thomas Transue
Cody Simmons
Audrey Wilkinson

Badger Technical Services

Joe Swintek

SeqAPASS v5.0

LaLone.Carlie@epa.gov

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