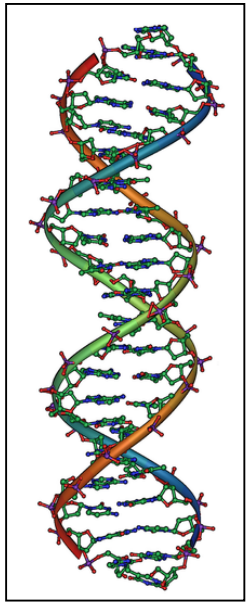
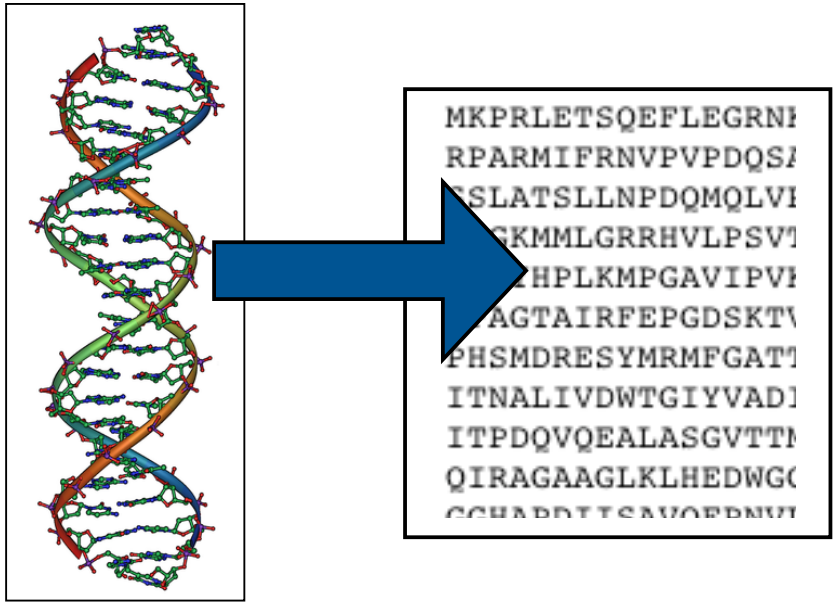
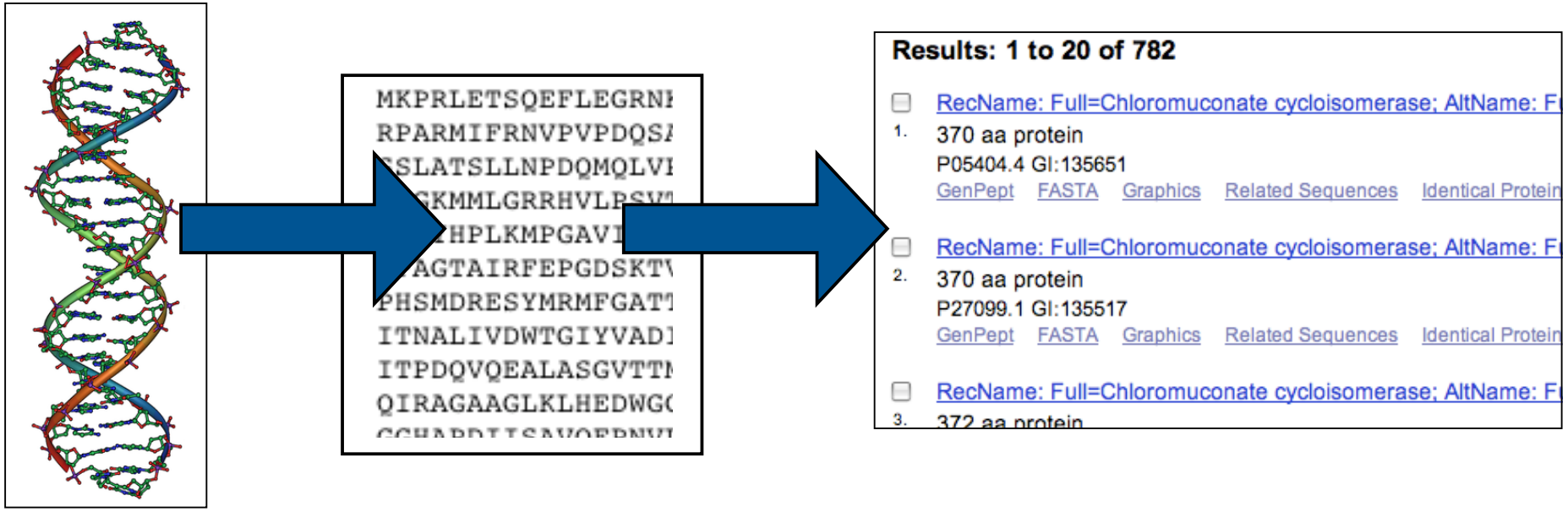
Annotation Error in Public Databases

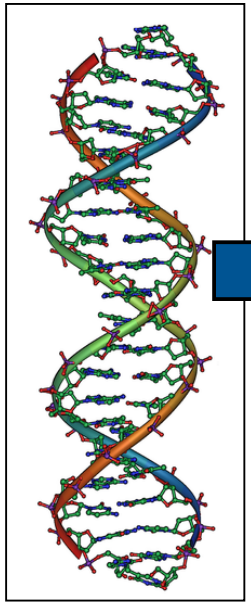
ALEXANDRA SCHNOES
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO
OCTOBER 25, 2010

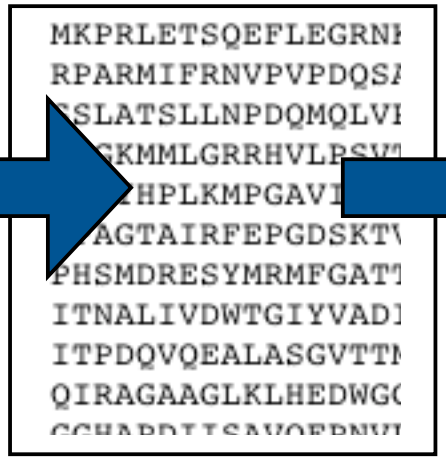
New genomes (and metagenomes) sequenced every day...

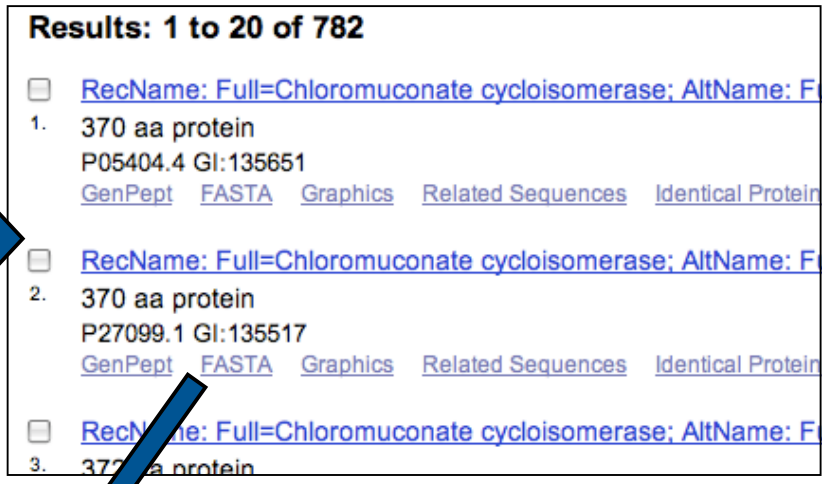


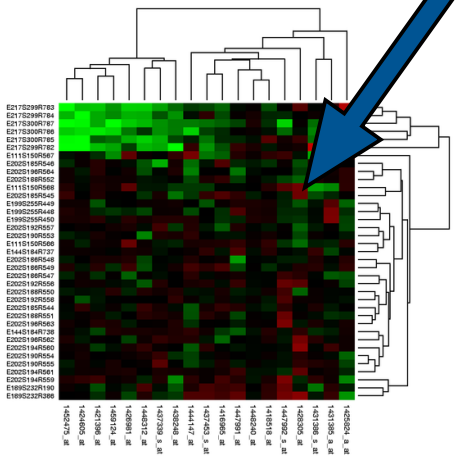

MKPRLETSQEFLEGRN
RPARMIFRNVPVPDQS
SLATSLLNPDQMQLV
KMMLGRRHVLPSV
HPLKMPGAVIPV
AGTAIRFEPGDSKT
PHSMDRESYMRMFGAT
ITNALIVDWTGIYVAD
ITPDQVQEALASGVTT
QIRAGAAGLKLHEDWG

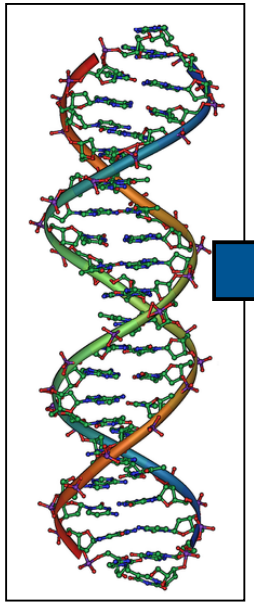

MKPRLETSQEFLEGRN
RPARMIFRNVPVPDQS
SLATSLLNPDQMQLV
KMMLGRRHVL
HPLKMPGAV
AGTAIRFEPGDSKT
PHSMDRESYMRMFGAT
ITNALIVDWTGIYVAD
ITPDQVQEALASGVTT
QIRAGAAGLKLHEDWG
Results: 1 to 20 of 782
RecName: Full=Chloromuconate cycloisomerase; AltName: F
1.
370 aa protein
P05404.4 GI:135651
GenPept FASTA Graphics Related Sequences Identical Protein
RecName: Full=Chloromuconate cycloisomerase; AltName: F
2.
370 aa protein
P27099.1 GI:135517
GenPept FASTA Graphics Related Sequences Identical Protein
RecName: Full=Chloromuconate cycloisomerase; AltName: F
3.
372 aa protein

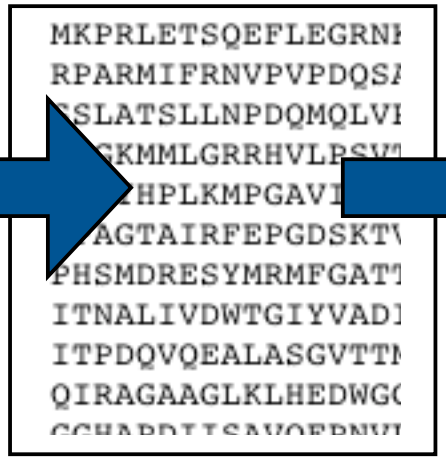


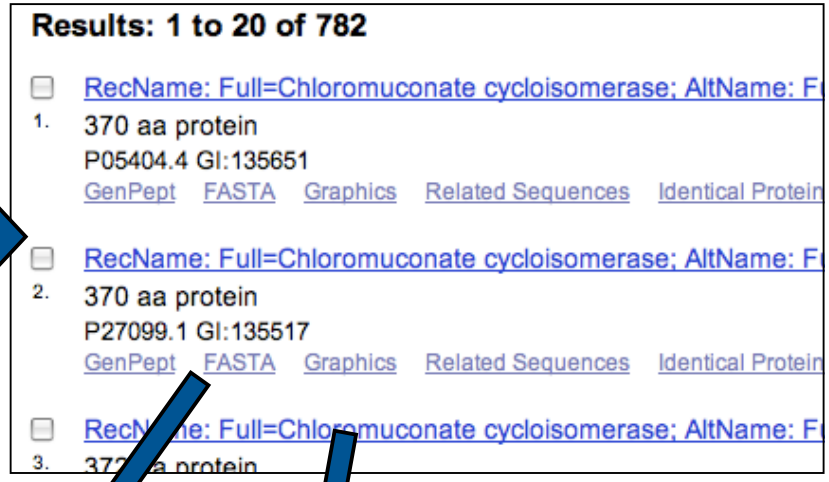

MKPRLETSQEFLEGRN
RPARMIFRNVPVPDQS
SLATSLLNPDQMQLV
KMMLGRRHVL
HPLKMPGAV
AGTAIRFEPGDSKT
PHSMDRESYMRMFGAT
ITNALIVDWTGIYVAD
ITPDQVQEALASGVTT
QIRAGAAGLKLHEDWG

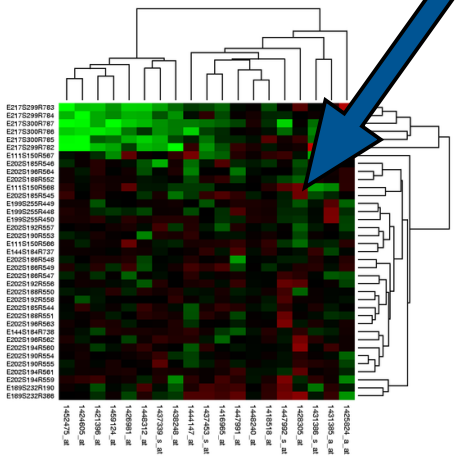

Results: 1 to 20 of 782
RecName: Full=Chloromuconate cycloisomerase; AltName: F
1. 370 aa protein
P05404.4 GI:135651
GenPept FASTA Graphics Related Sequences Identical Protein
RecName: Full=Chloromuconate cycloisomerase; AltName: F
2. 370 aa protein
P27099.1 GI:135517
GenPept FASTA Graphics Related Sequences Identical Protein
RecName: Full=Chloromuconate cycloisomerase; AltName: F
3.



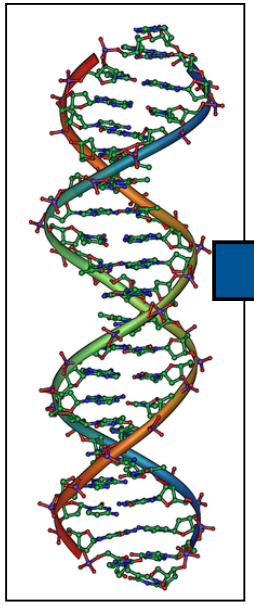



MKPRLETSQEFLEGRN
RPARMIFRNVPVPDQS
SLATSLLNPDQMQLV
KMMLGRRHVL
HPLKMPGAV
AGTAIRFEPGDSKT
PHSMDRESYMRMFGAT
ITNALIVDWTGIYVAD
ITPDQVQEALASGVTT
QIRAGAAGLKLHEDWG


Results: 1 to 20 of 782
RecName: Full=Chloromuconate cycloisomerase; AltName: F
1. 370 aa protein
P05404.4 GI:135651
GenPept FASTA Graphics Related Sequences Identical Protein
RecName: Full=Chloromuconate cycloisomerase; AltName: F
2. 370 aa protein
P27099.1 GI:135517
GenPept FASTA Graphics Related Sequences Identical Protein
RecName: Full=Chloromuconate cycloisomerase; AltName: F
3.



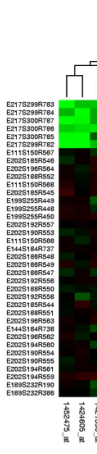

RESEARCH ARTICLE
Sequencing and Analysis
Neanderthal Genomic DNA



MKPRLETSQEFLEGRN
RPARMIFRNVPVPDQS
SLATSLLNPDQMQLV
KMMLGRRHVL
HPLKMPGAV
AGTAIRFEPGDSKT
PHSMDRESYMRMFGAT
ITNALIVDWTGIYVAD
ITPDQVQEALASGVTT
QIRAGAAGLKLHEDWG

Results: 1 to 20 of 782

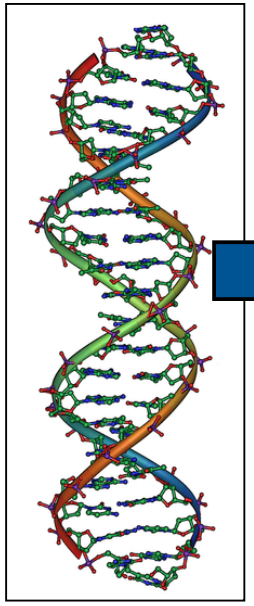
1. RecName: Full=Chloromuconate cycloisomerase; AltName: F
   370 aa protein
   P05404.4 GI:135651
   GenPept FASTA Graphics Related Sequences Identical Protein
2. RecName: Full=Chloromuconate cycloisomerase; AltName: F
   370 aa protein
   P27099.1 GI:135517
   GenPept FASTA Graphics Related Sequences Identical Protein
3. RecName: Full=Chloromuconate cycloisomerase; AltName: F
   372 aa protein



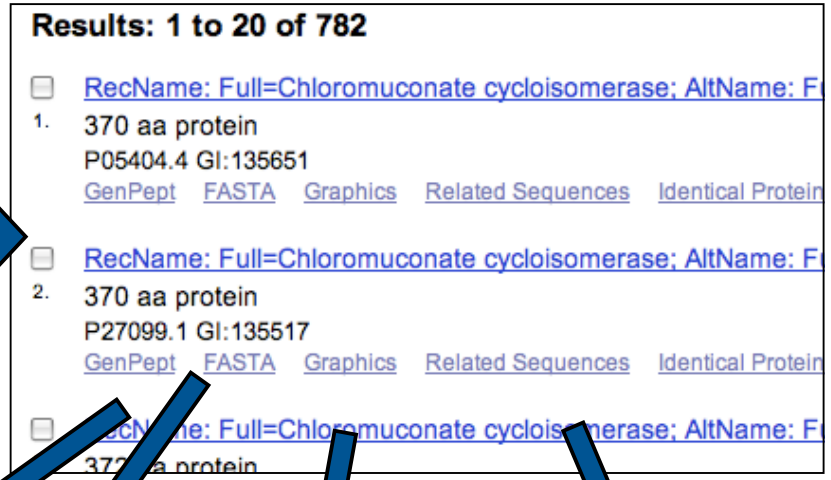
RESEARCH ARTICLE

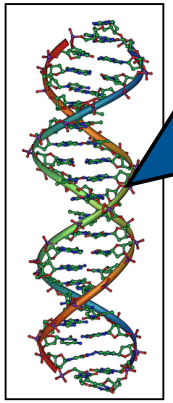
Sequencing and Analysis of
Neanderthal Genomic DNA

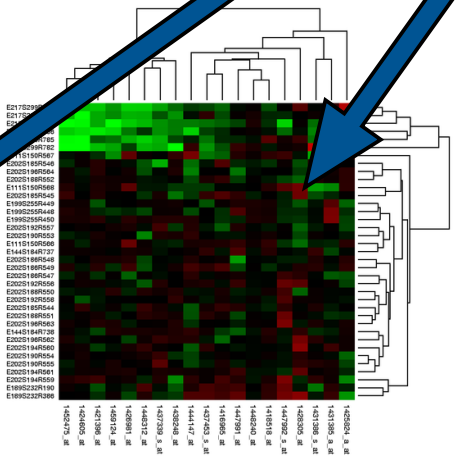
1. RecName: Full=Chlor
   370 aa protein
   P05404.4 GI:135651
   GenPept FASTA Gra
2. RecName: Full=Chlor
   370 aa protein
   P27099.1 GI:135517
   GenPept FASTA Gra
3. RecName: Full=Chlor
   372 aa protein



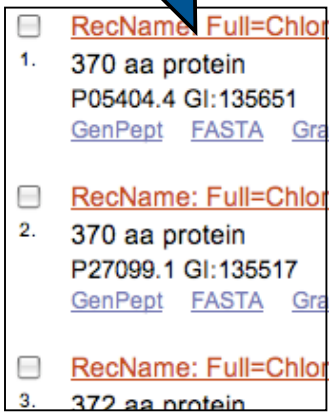

MKPRLETSQEFLEGRN
RPARMIFRNVPVPDQS
SLATSLLNPDQMQLV
KMMLGRRHVL
HPLKMPGAV
AGTAIRFEPGDSKT
PHSMDRESYMRMFGAT
ITNALIVDWTGIYVAD
ITPDQVQEALASGVTT
QIRAGAAGLKLHEDWG

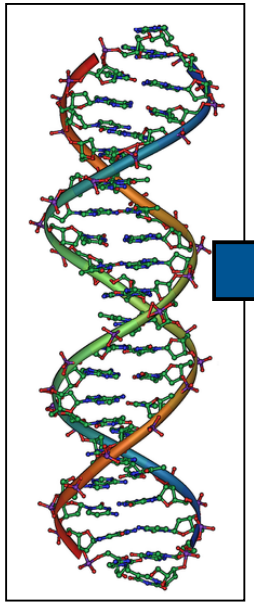

Results: 1 to 20 of 782
RecName: Full=Chloromuconate cycloisomerase; AltName: F
1.
370 aa protein
P05404.4 GI:135651
GenPept FASTA Graphics Related Sequences Identical Protein
RecName: Full=Chloromuconate cycloisomerase; AltName: F
2.
370 aa protein
P27099.1 GI:135517
GenPept FASTA Graphics Related Sequences Identical Protein

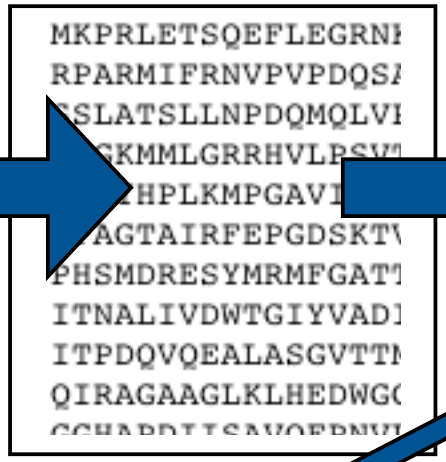


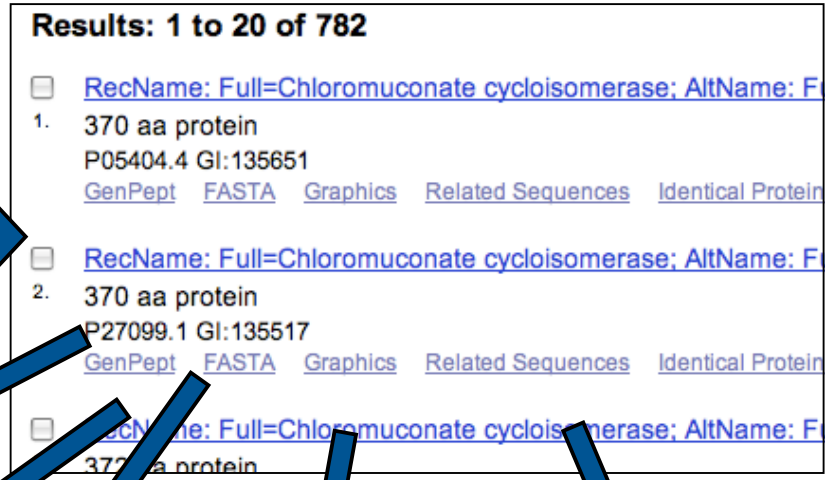


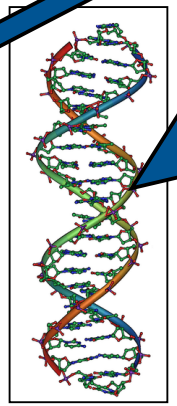

RESEARCH ARTICLE
Sequencing and Analysis
Neanderthal Genomic DNA

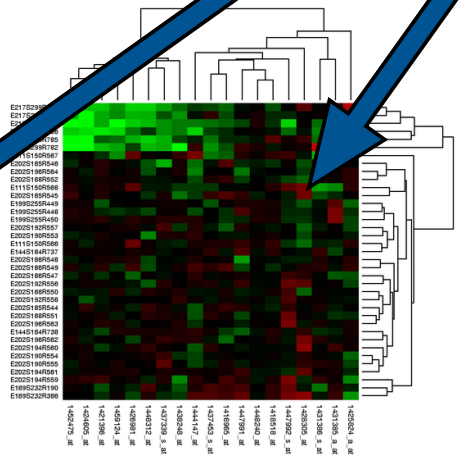

RecName: Full=Chlor
1.
370 aa protein
P05404.4 GI:135651
GenPept FASTA Gra
RecName: Full=Chlor
2.
370 aa protein
P27099.1 GI:135517
GenPept FASTA Gra
RecName: Full=Chlor
3.
372 aa protein



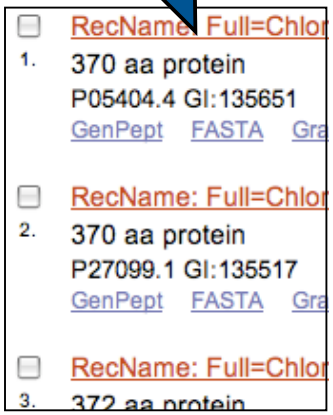

MKPRLETSQEFLEGRN
RPARMIFRNVPVPDQS
SLATSLLNPDQMQLV
KMMLGRRHVL
HPLKMPGAV
AGTAIRFEPGDSKT
PHSMDRESYMRMFGAT
ITNALIVDWTGIYVAD
ITPDQVQEALASGVTT
QIRAGAAGLKLHEDWG

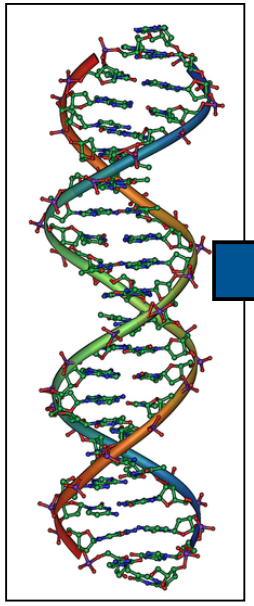

Results: 1 to 20 of 782
RecName: Full=Chloromuconate cycloisomerase; AltName: F
1.
370 aa protein
P05404.4 GI:135651
GenPept FASTA Graphics Related Sequences Identical Protein
RecName: Full=Chloromuconate cycloisomerase; AltName: F
2.
370 aa protein
P27099.1 GI:135517
GenPept FASTA Graphics Related Sequences Identical Protein

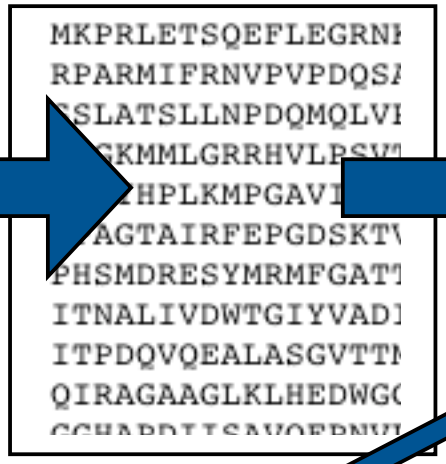


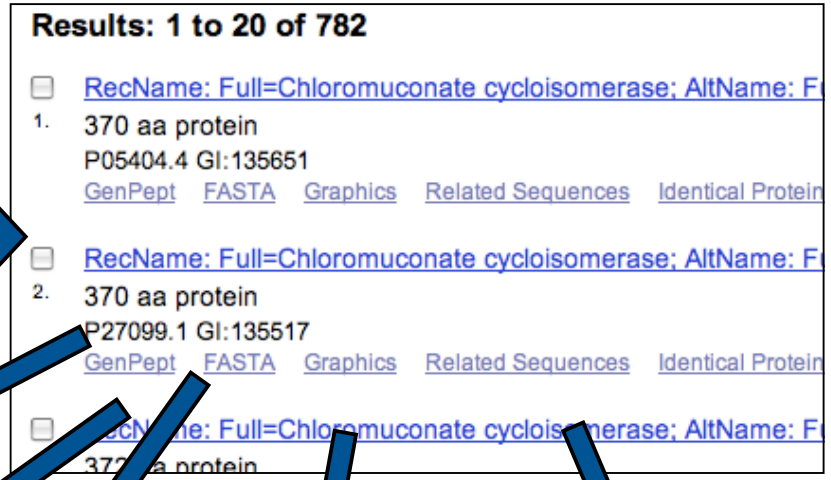


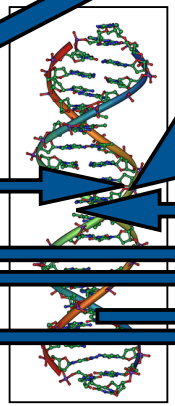

RESEARCH ARTICLE
Sequencing and Analysis of
Neanderthal Genomic DNA

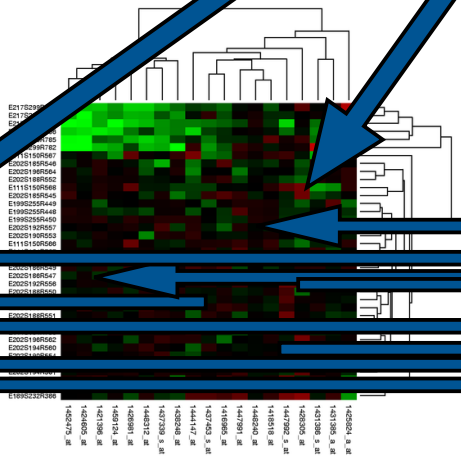

RecName: Full=Chlor
1.
370 aa protein
P05404.4 GI:135651
GenPept FASTA Gra
RecName: Full=Chlor
2.
370 aa protein
P27099.1 GI:135517
GenPept FASTA Gra
RecName: Full=Chlor
3.
372 aa protein

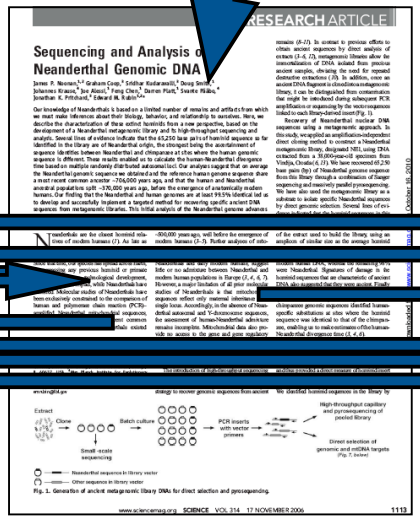


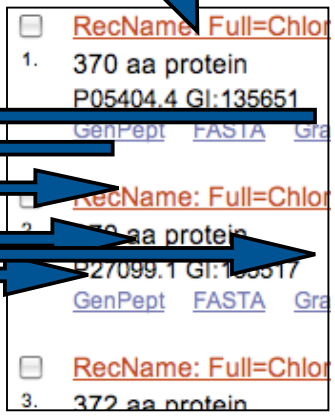

MKPRLETSQEFLEGRNI
RPARMIFRNVPVPDQSA
SLATSLLNPDQMQLVI
KMMLGRRHVLPSV
HPLKMPGAVI
AGTAIRFEPGDSKTV
PHSMDRESYMRMFGATT
ITNALIVDWTGIYVADI
ITPDQVQEALASGVTTN
QIRAGAAGLKLHEDWGC


Results: 1 to 20 of 782
RecName: Full=Chloromuconate cycloisomerase; AltName: F
1. 370 aa protein
P05404.4 GI:135651
GenPept FASTA Graphics Related Sequences Identical Protein
RecName: Full=Chloromuconate cycloisomerase; AltName: F
2. 370 aa protein
P27099.1 GI:135517
GenPept FASTA Graphics Related Sequences Identical Protein






RESEARCH ARTICLE
Sequencing and Analysis of
Neanderthal Genomic DNA


RecName: Full=Chlor
1. 370 aa protein
P05404.4 GI:135651
GenPept FASTA Gra
RecName: Full=Chlor
RecName: Full=Chlor
3. 372 aa protein

Computational Function Prediction Needed

Total Sequences

Characterized Sequences

What about the error that results from large scale function prediction?

Our focus: commonly used protein sequence databases

How prevalent is misannotation in common sequence databases?

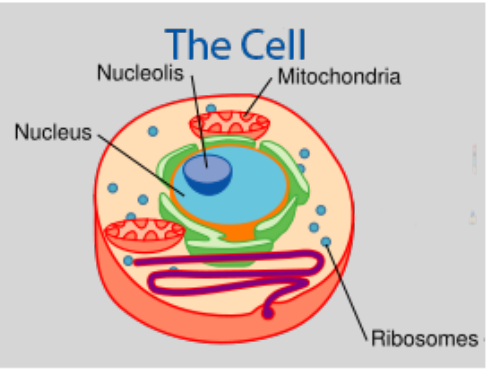
What can we learn about these annotation errors and annotation in general?

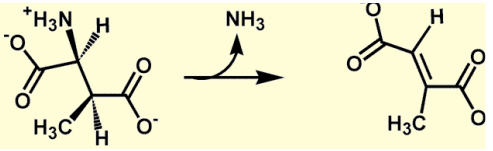
What is ‘function’?

Many Possible Definitions

- Biological role
- Enzymatic activity
- Protein localization
- Protein interactions
- Protein expression
- Temporal activity
- Post-translational modifications
- Structural domain
- Sequence motif
- Structural motif
- Binding sites
- Functionally important residues
- Genomic context
- Metabolic pathway

Phenotype





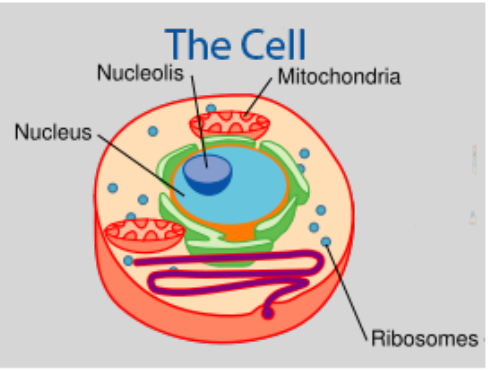
Enzymatic Reaction

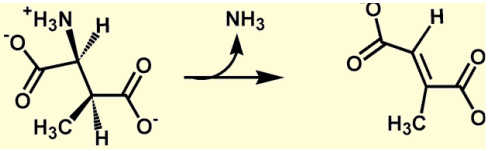
What is ‘function’?

Many Possible Definitions

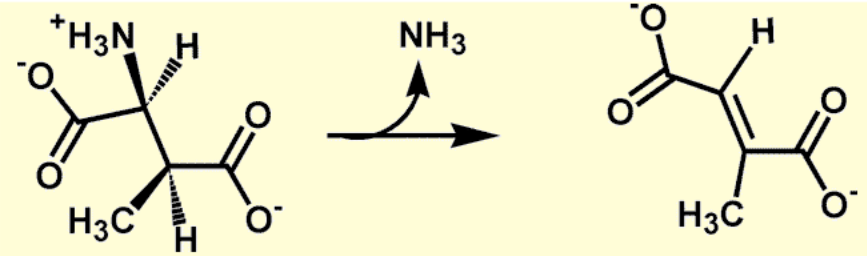
- Biological role
- Enzymatic activity
- Protein localization
- Protein interactions
- Protein expression
- Temporal activity
- Post-translational modifications
- Structural domain
- Sequence motif
- Structural motif
- Binding sites
- Functionally important residues
- Genomic context
- Metabolic pathway

Phenotype

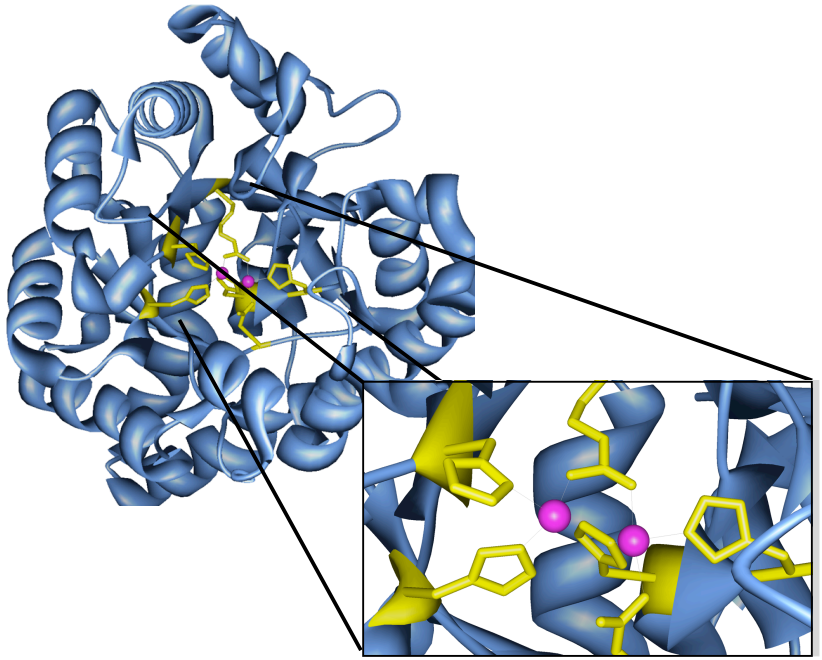




Enzymatic Reaction



- Concrete definition of function
  - Substrate
  - Product
  - Chemical conversion

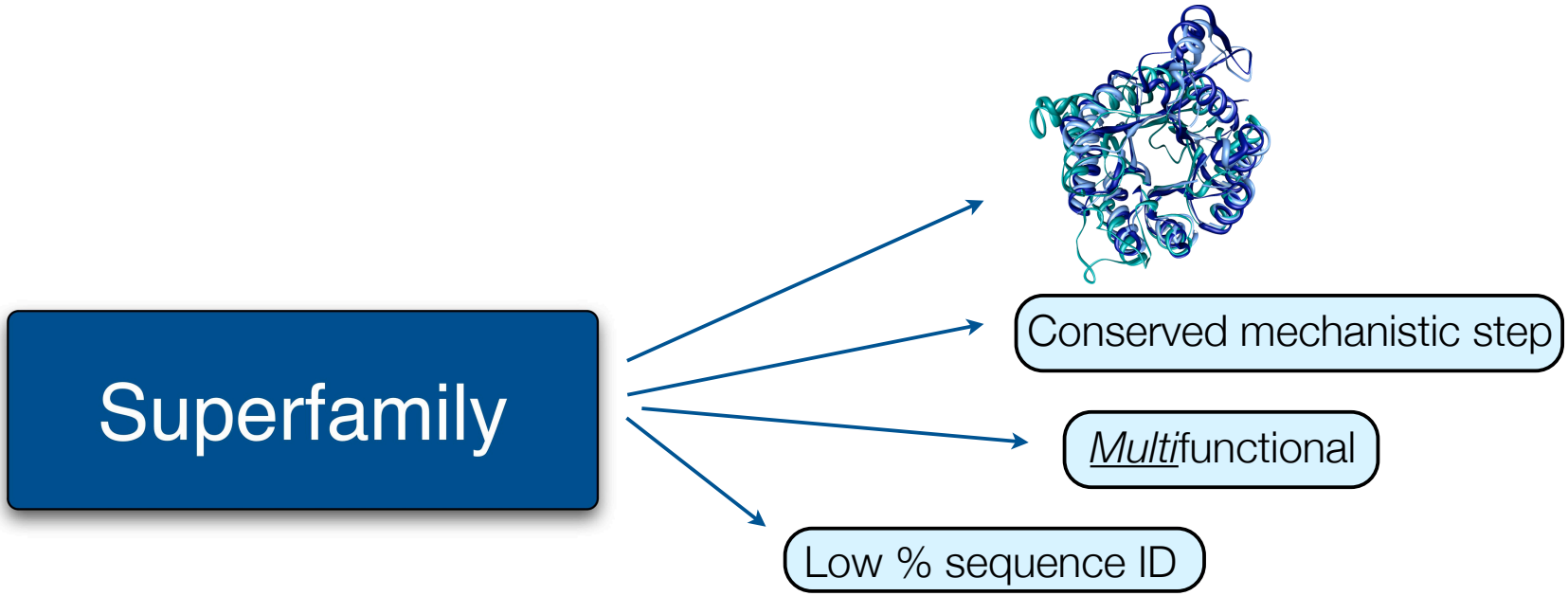


- Function can be mapped to specific residues

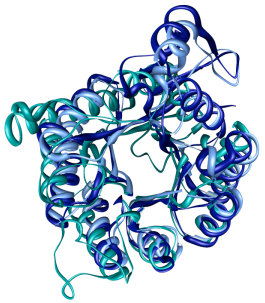
Why use enzymes?

Functionally Diverse Enzyme Superfamilies

Functionally Diverse Enzyme Superfamilies



Functionally Diverse Enzyme Superfamilies



Superfamily

Conserved mechanistic step

*Multi*functional

Low % sequence ID

*Mono*functional

Family Specific Residues

% ID within families
>
% ID between families

Family

Family

Family

...

What is needed for the misannotation analysis?

Gold Standard Sequence Set

Requirements

- Organized hierarchy & data
  - Superfamily definitions
  - Family definitions
  - Sequences
  - Sequence alignments
  - Statistical models
- Functions are experimentally characterized
- Understand functional mechanism
  - Structure
  - Active site
  - Functionally important residues
- Large set

What is needed for the misannotation analysis?

Gold Standard Sequence Set

Requirements

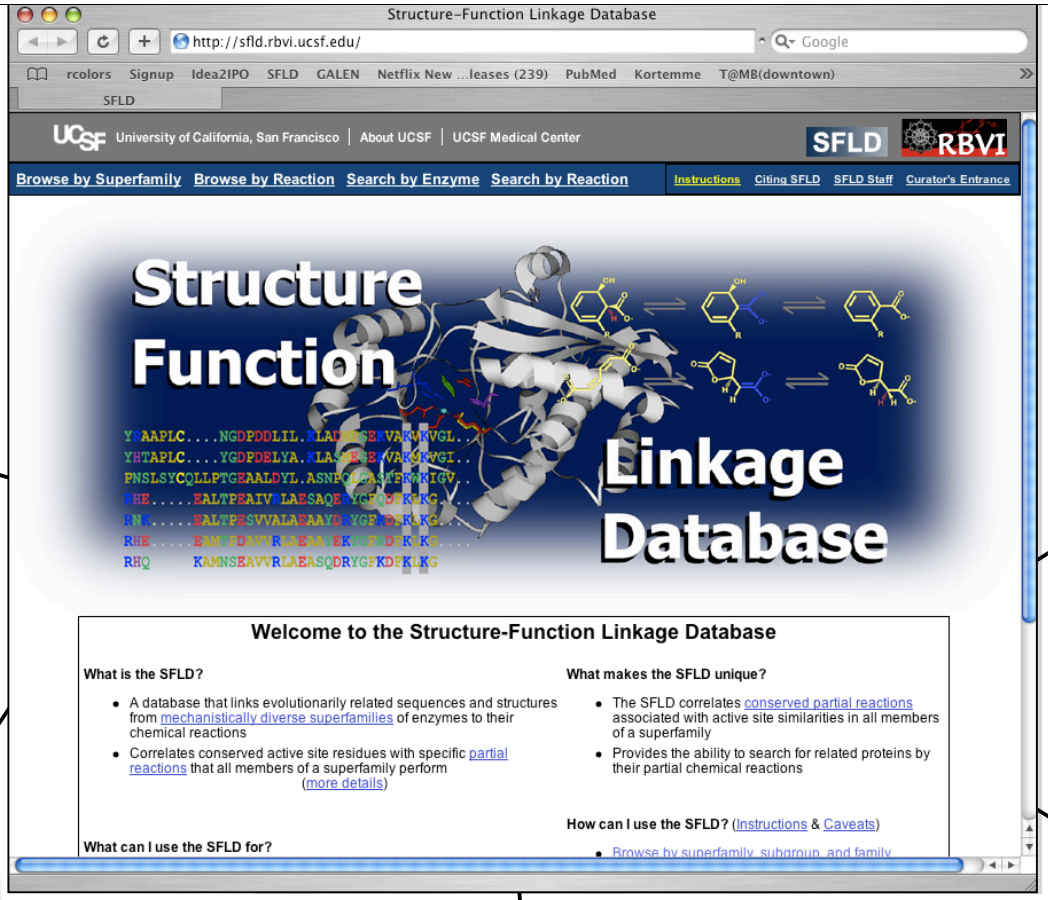
- Organized hierarchy & data
  - Superfamily definitions
  - Family definitions
  - Sequences
  - Sequence alignments
  - Statistical models
- Functions are experimentally characterized
- Understand functional mechanism
  - Structure
  - Active site
  - Functionally important residues
- Large set

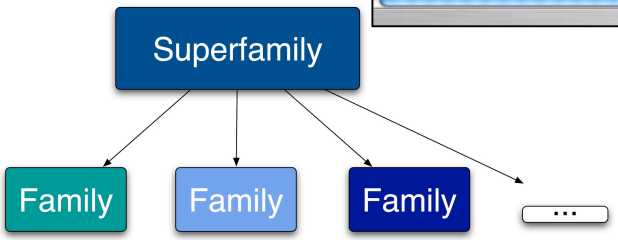
— 6 Superfamilies
— 5 Structural folds
— 37 Families
—5/6 E.C. categories

Genome Biol. 2006;7(1):R8.

Evidence Codes

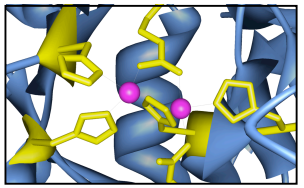
Sequence Models (HMMs)





Hierarchically Organized

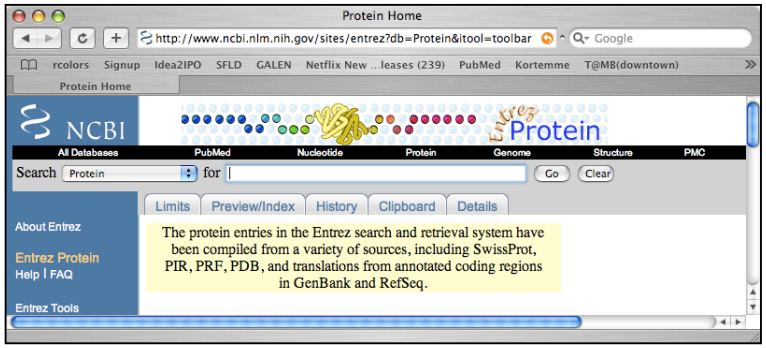
Hand-Curated Sequence Alignments



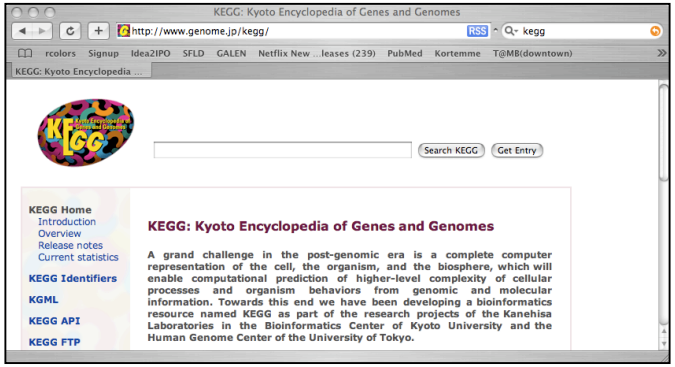
Functionally Important Residues

Gold Standard Sequence Set | *sfld.rbvi.ucsf.edu*

Data Source: Commonly Used Sequence Databases



NCBI
Automated
Large



KEGG
Automated

TrEMBL
Automated
Large

Swiss-Prot
Curated
Small

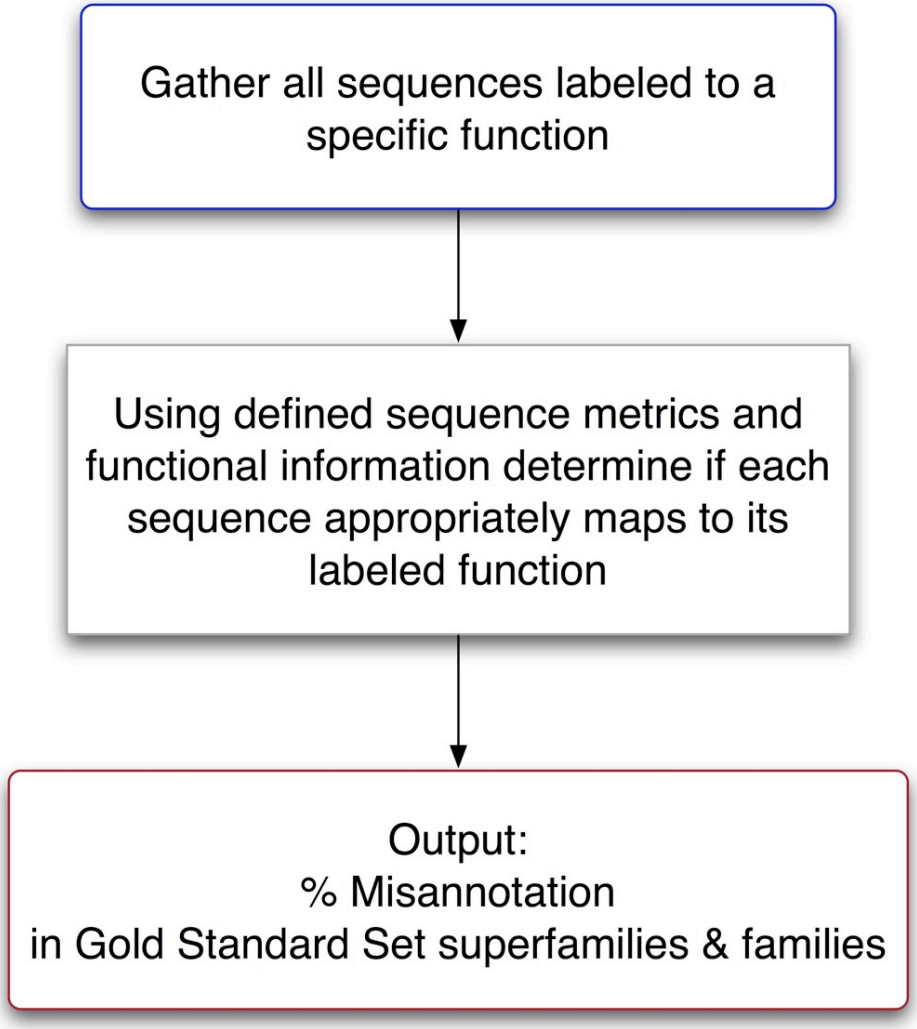
Analysis Question

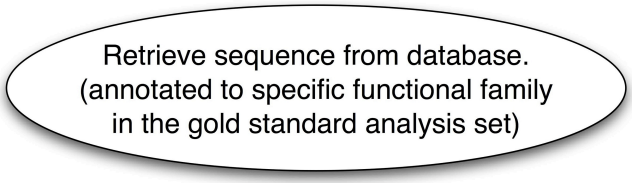
Given:

A protein sequence annotated to a specific enzyme function

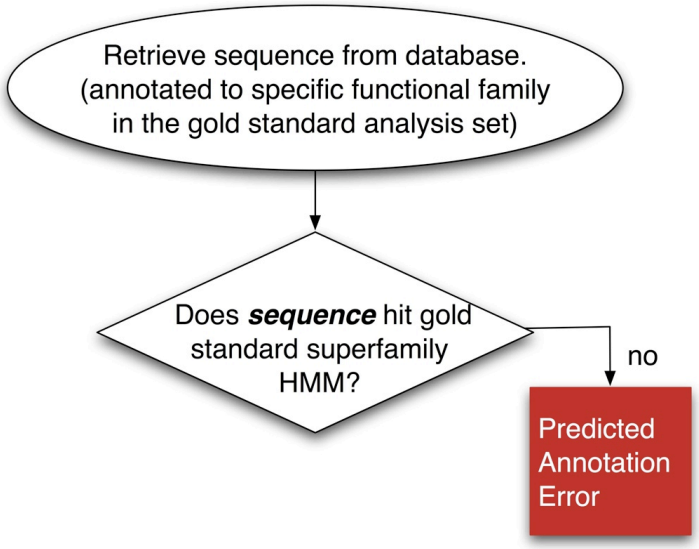
Is that annotation correct?

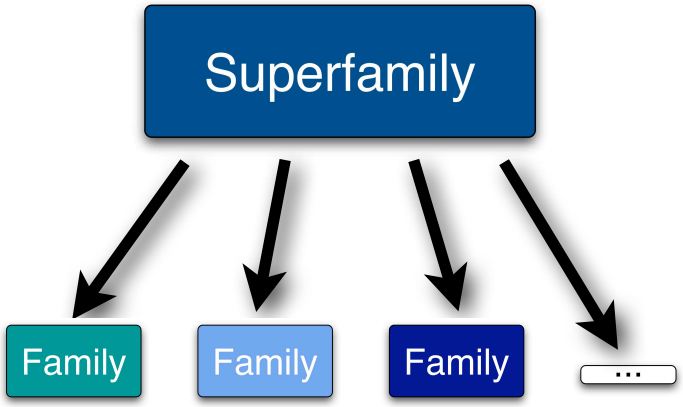
General Process

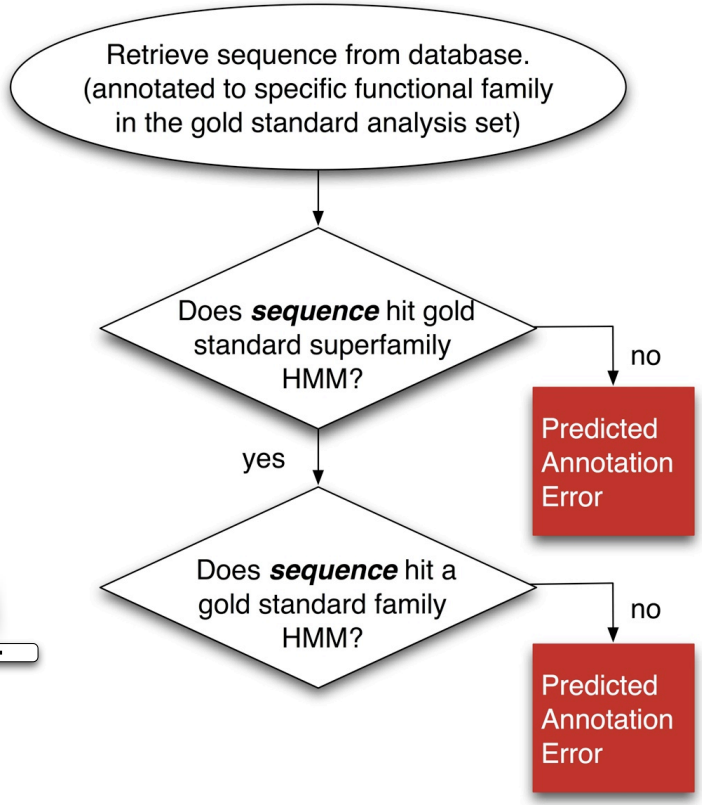


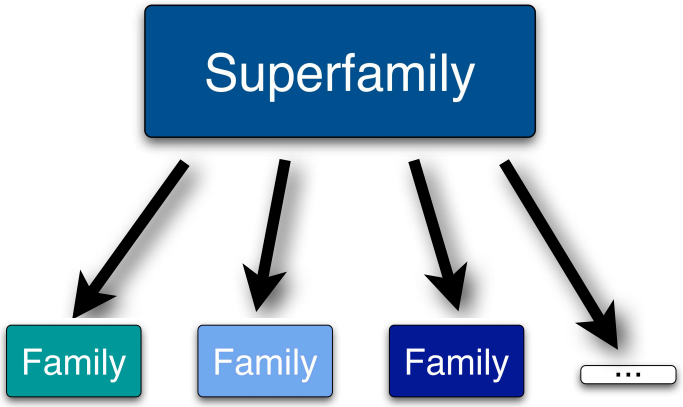

Retrieve sequence from database.
(annotated to specific functional family
in the gold standard analysis set)

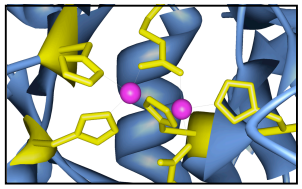
Superfamily

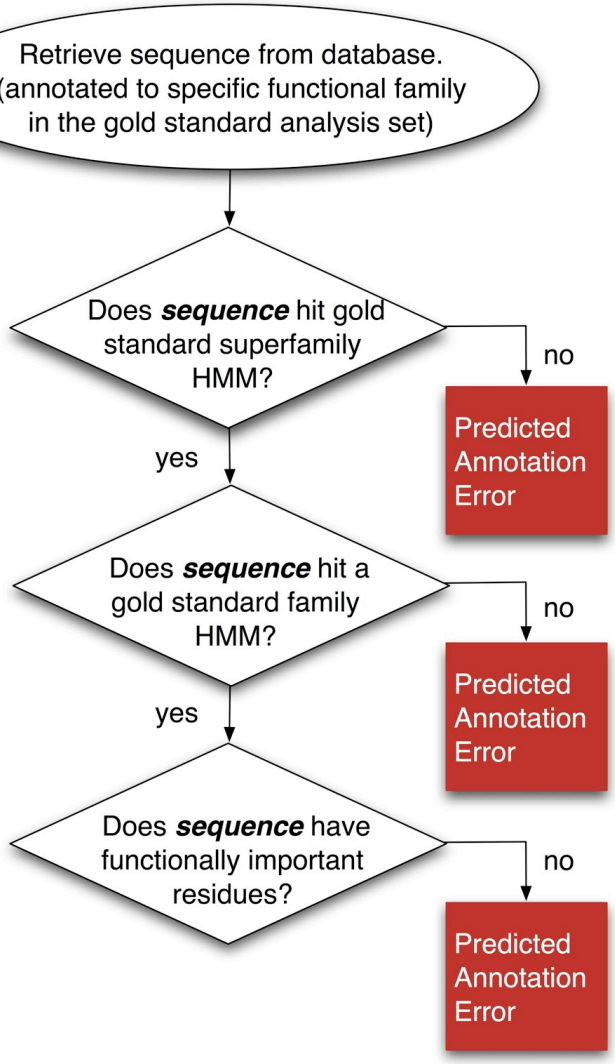



Superfamily
Family
Family
Family
...

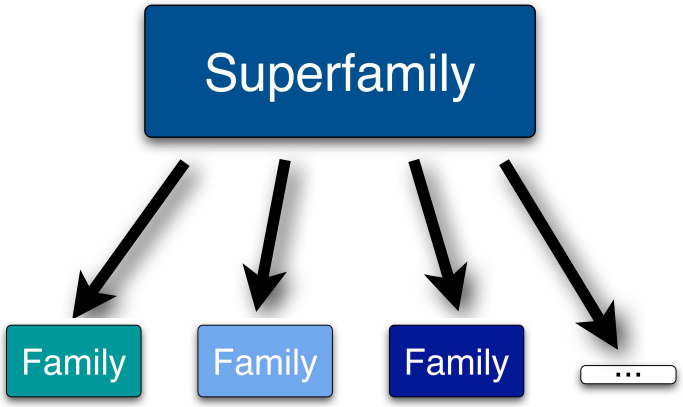

Retrieve sequence from database.
(annotated to specific functional family
in the gold standard analysis set)
Does sequence hit gold standard superfamily HMM?
no
Predicted Annotation Error
yes
Does sequence hit a gold standard family HMM?
no
Predicted Annotation Error

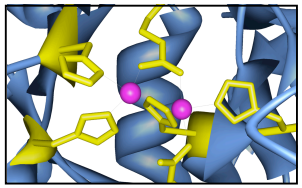

Superfamily
Family
Family
Family
...

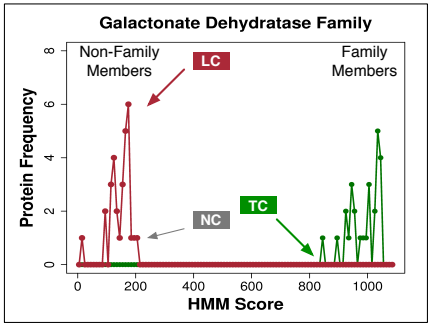


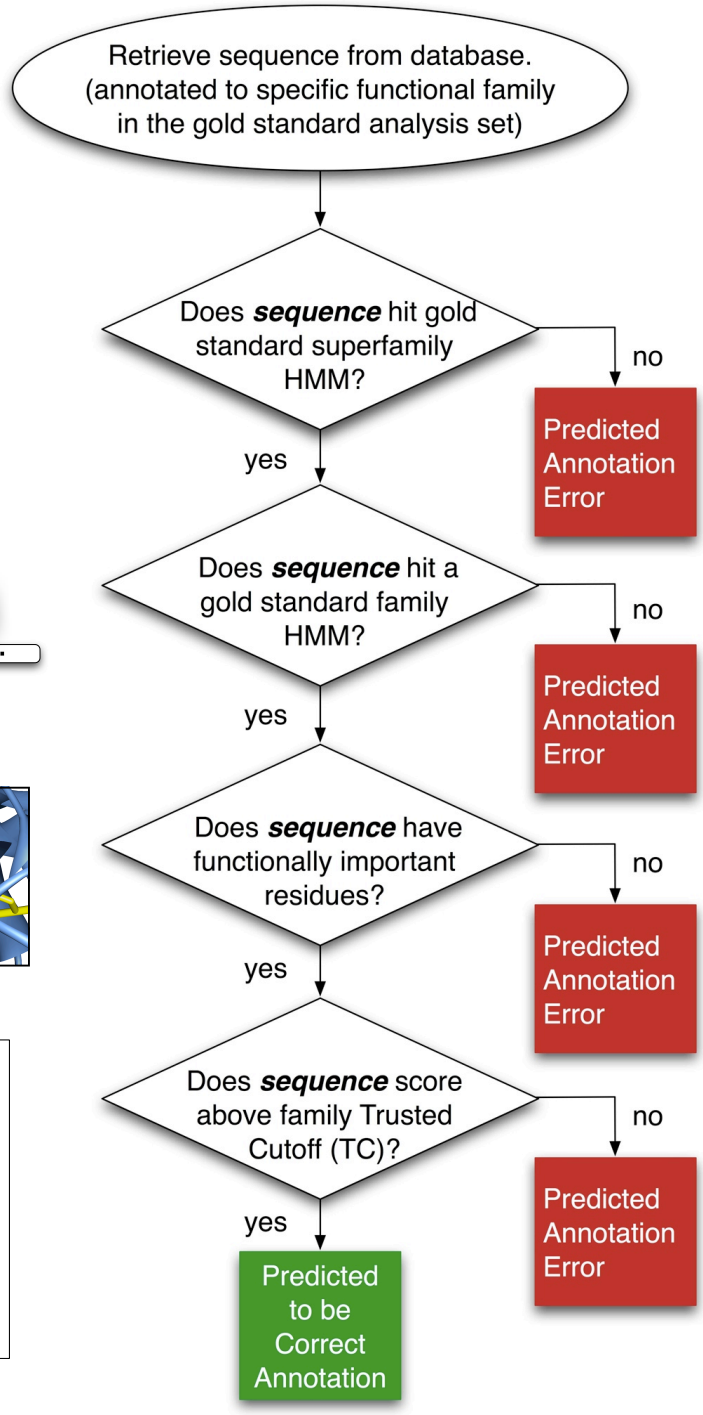

Retrieve sequence from database. (annotated to specific functional family in the gold standard analysis set)
Does sequence hit gold standard superfamily HMM?
no
Predicted Annotation Error
yes
Does sequence hit a gold standard family HMM?
no
Predicted Annotation Error
yes
Does sequence have functionally important residues?
no
Predicted Annotation Error


300
310
320
330

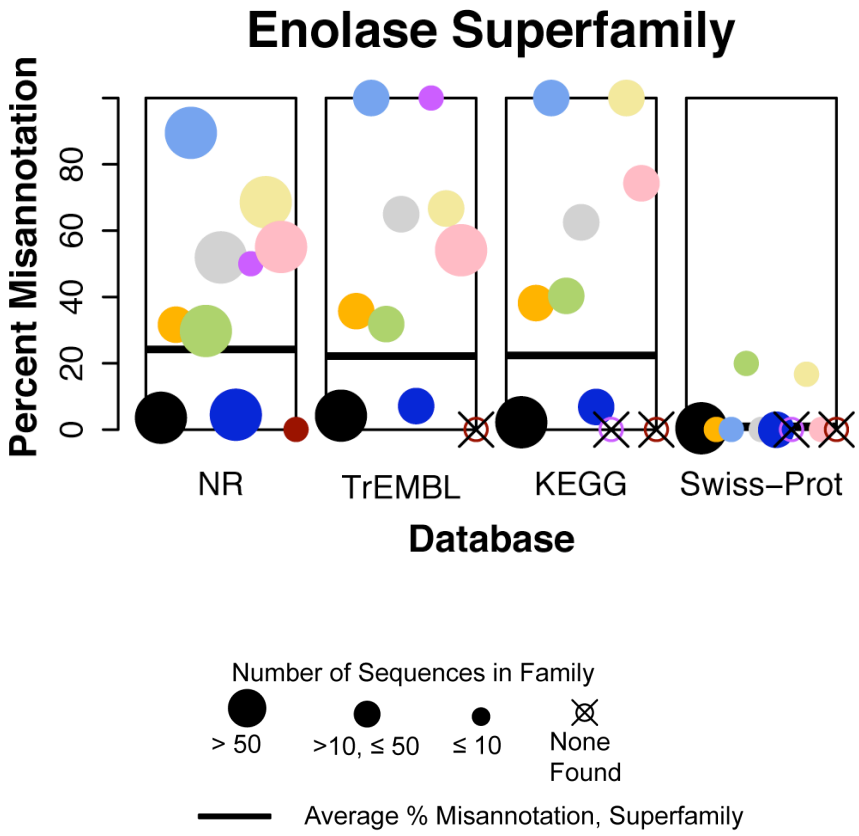

Superfamily
Family
Family
Family
...

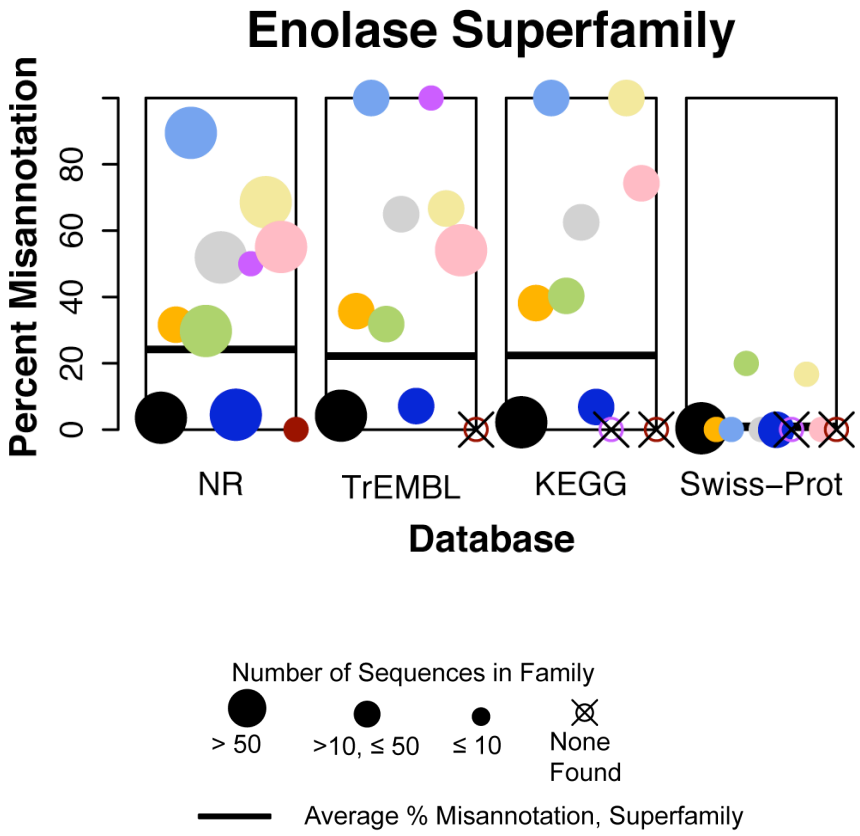



Galactonate Dehydratase Family
Non-Family Members
LC
Family Members
NC
TC
Protein Frequency
0
2
4
6
8
0
200
400
600
800
1000
HMM Score


Retrieve sequence from database. (annotated to specific functional family in the gold standard analysis set)
Does sequence hit gold standard superfamily HMM?
no
Predicted Annotation Error
yes
Does sequence hit a gold standard family HMM?
no
Predicted Annotation Error
yes
Does sequence have functionally important residues?
no
Predicted Annotation Error
yes
Does sequence score above family Trusted Cutoff (TC)?
no
Predicted Annotation Error
yes
Predicted to be Correct Annotation


300
310
320
330

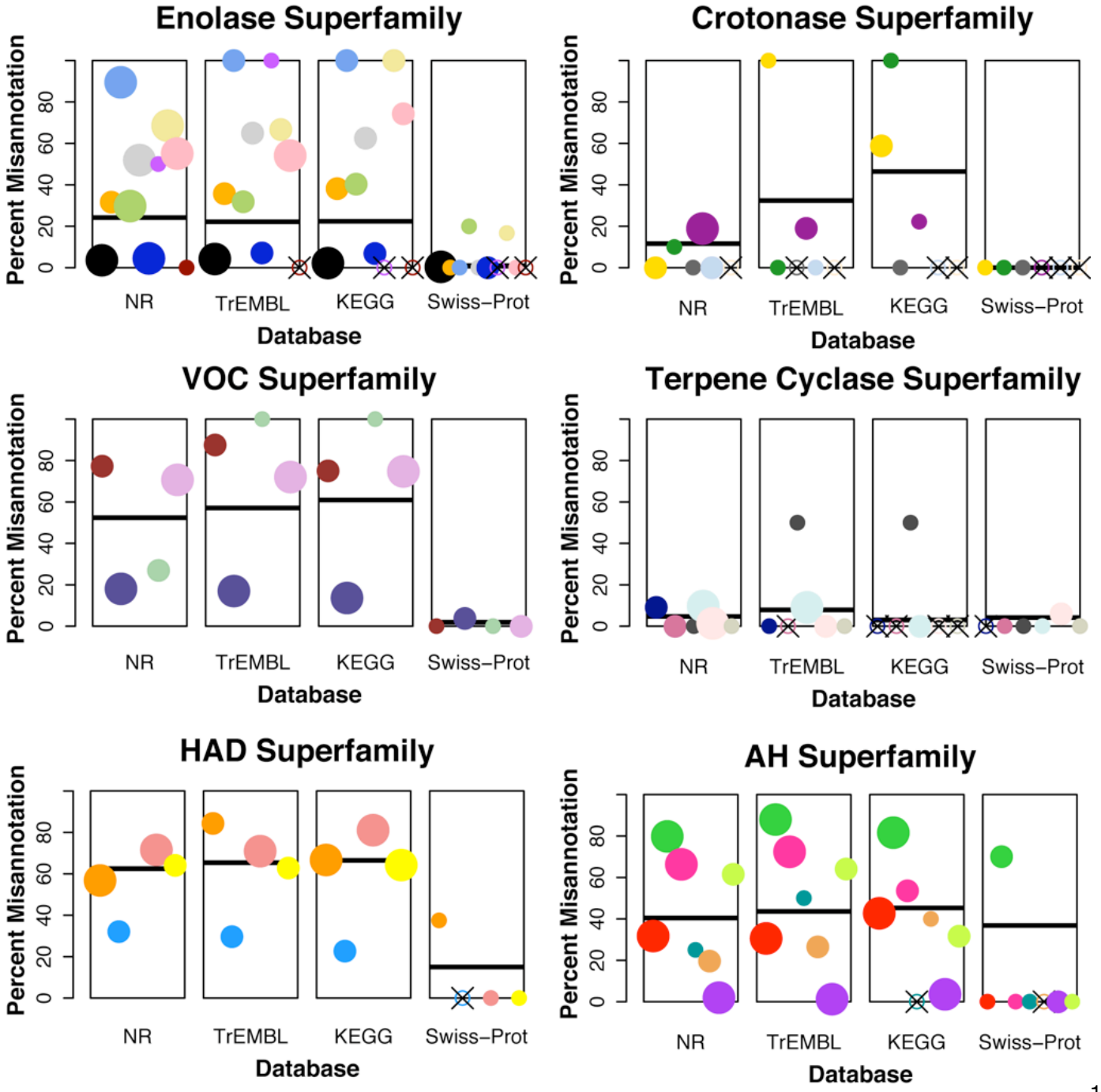

Enolase Superfamily
Percent Misannotation
0
20
40
60
80
NR
TrEMBL
KEGG
Swiss−Prot
Database
Number of Sequences in Family
> 50
>10, ≤ 50
≤ 10
None Found
Average % Misannotation, Superfamily



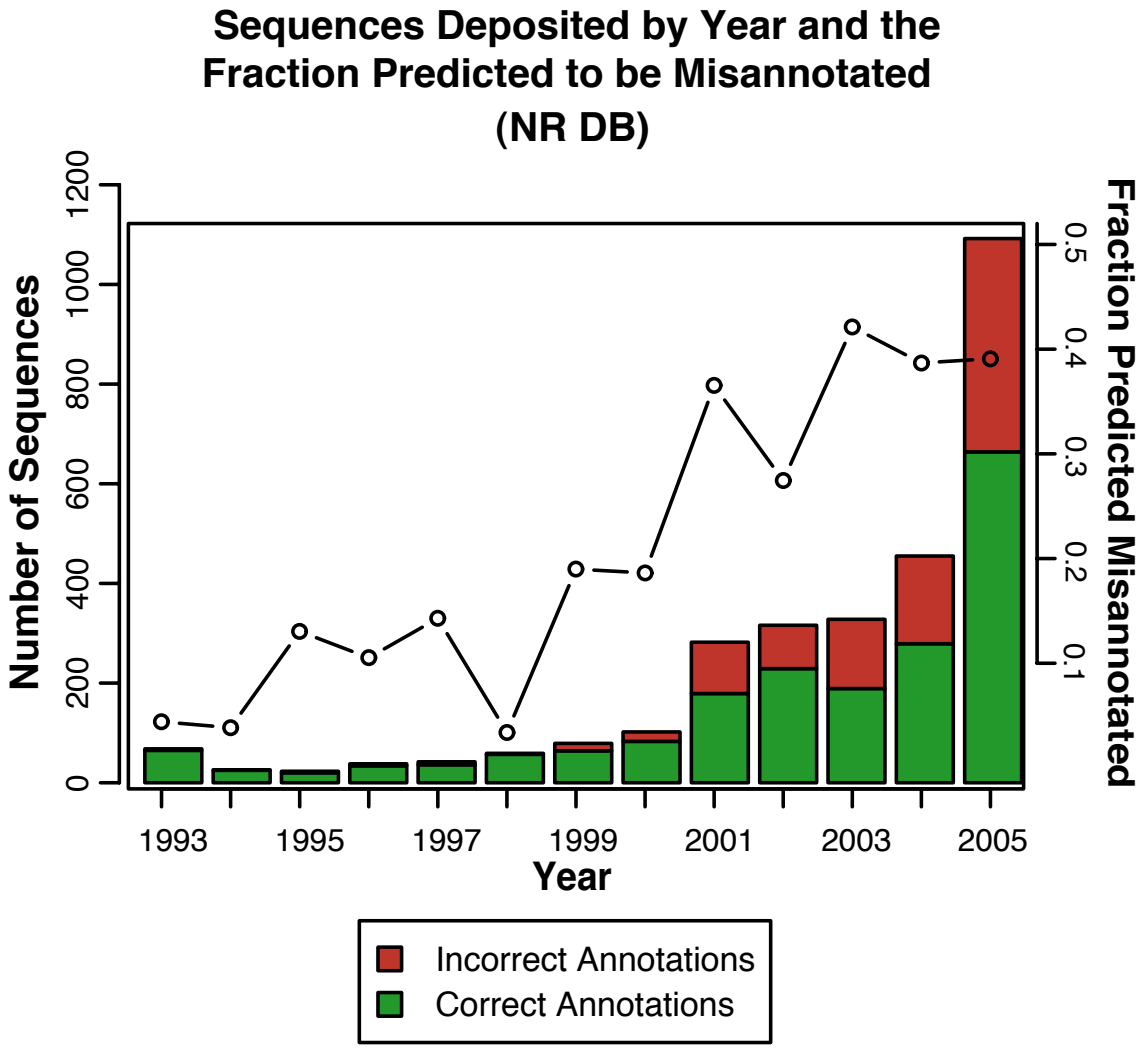
Superfamily	Family	E.C. No.	Family Color
Enolase	Enolase	4.2.1.11	
	Galactonate dehydratase	4.2.1.6	
	Mandelate racemase	5.1.2.2	
	Glucarate dehydratase	4.2.1.40	
	Methylaspartate ammonia-lyase	4.3.1.2	
	ortho-succinyl benzoate synthase	4.2.1.113	
	Dipeptide epimerase	—	
	Chloromuconate cycloisomerase	5.5.1.7	
	Muconate cycloisomerase	5.5.1.1	
	L-fuconate dehydratase	4.2.1.68	
Crotonase	Dodecenoyl-CoA delta-isomerase (mitochondrial)	5.3.3.8	
	Delta(3,5)-delta(2,4)-dienoyl-CoA isomerase	—	
	Methylmalonyl-CoA decarboxylase	4.1.1.41	
	3-Hydroxyisobutyryl-CoA hydrolase	3.1.2.4	
	4-Chlorobenzoate dehalogenase	3.8.1.7	
	1,4-Dihydroxy-2-napthoyl-CoA synthase	—	
Vicinal Oxygen Chelate (VOC)	Methylmalonyl-CoA epimerase	5.1.99.1	
	4-Hydroxyphenylpyruvate dioxygenase	1.13.11.27	
	FosA	2.5.1.18	
	Glyoxalase I	4.4.1.5	
Terpene Cyclase	5-Epi-aristolochene synthase	—	
	Bornyl diphosphate synthase	5.5.1.8	
	Pentalenene synthase	4.2.3.7	
	Squalene-hopene synthase	5.4.99.17	
	Trichodiene synthase	4.2.3.6	
	Aristolochene synthase	4.2.3.9	
Haloacid Dehalogenase (HAD)	Deoxy-D-mannose-octulosonate 8-phosphate phosphatase	3.1.3.45	
	Phosphonoacetaldehyde hydrolase	3.11.1.1	
	2-Haloacid dehalogenase	3.8.1.2	
	Beta-phosphoglucomutase	5.4.2.6	
Amidohydrolase (AH)	Cytosine deaminase	3.5.4.1	
	Adenosine deaminase	3.5.4.4	
	N-acyl-D-amino-acid deacyclase	3.5.1.81	
	L-hydantoinase	3.5.2.2	
	D-hydantoinase	3.5.2.2	
	Urease	3.5.1.5	
	Isoaspartyl dipeptidase	—	

Variable percent misannotation

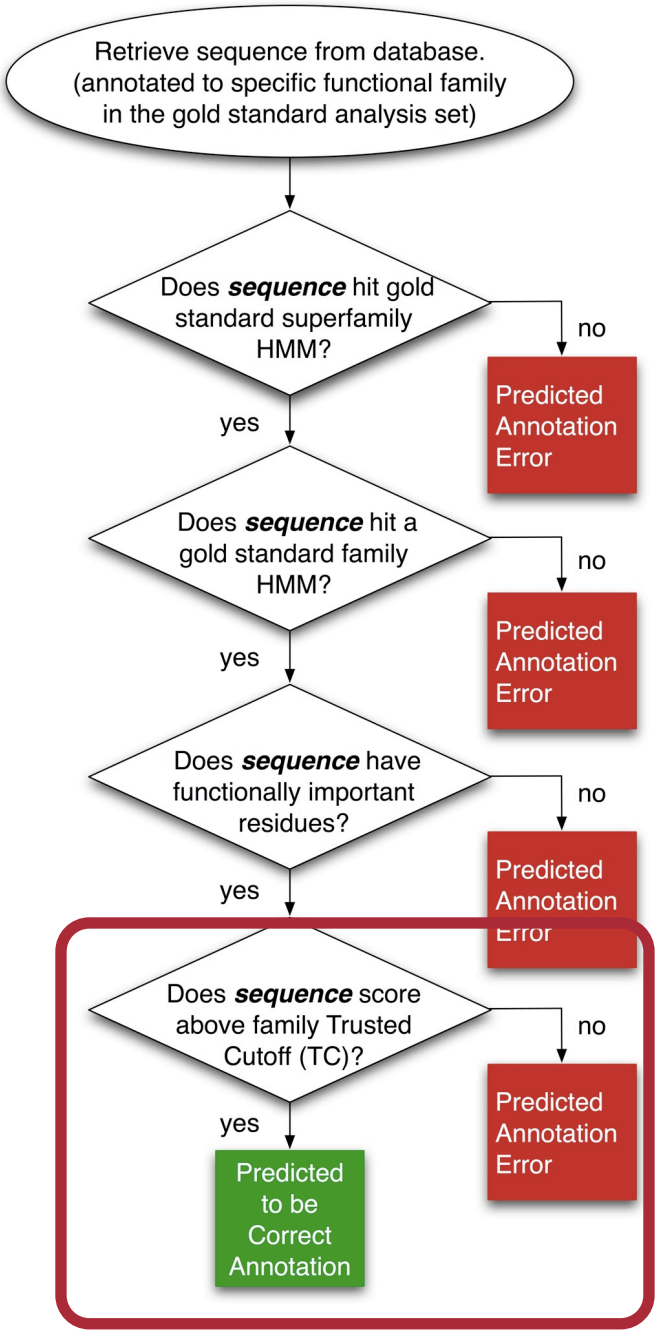
Manually curated Swiss-Prot is most accurate



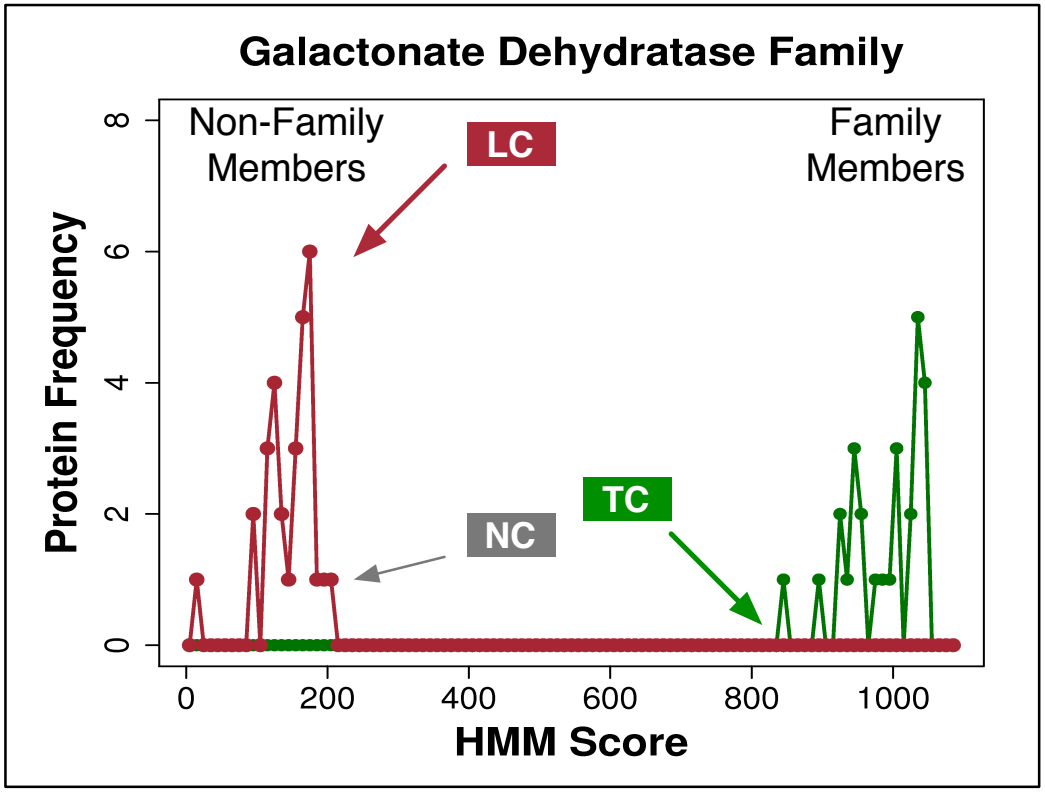
Misannotation Problem is Getting Worse



What are the characteristics of these misannotations?



Sensitivity to threshold change

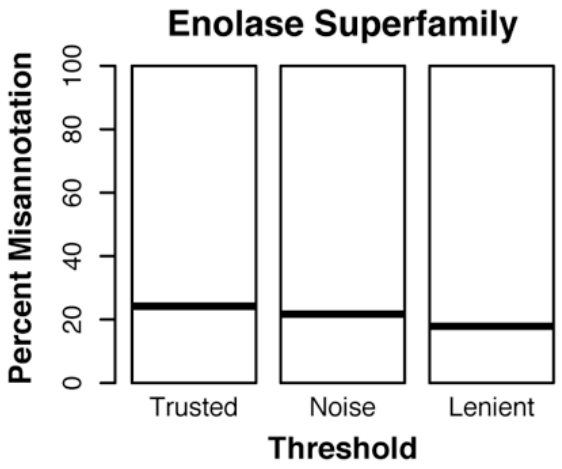


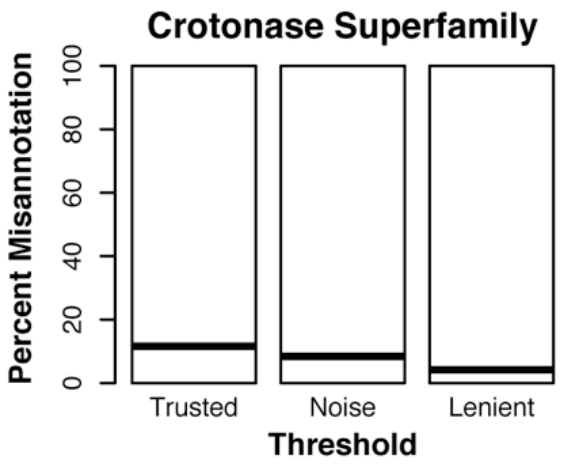
TC — Trusted Cutoff

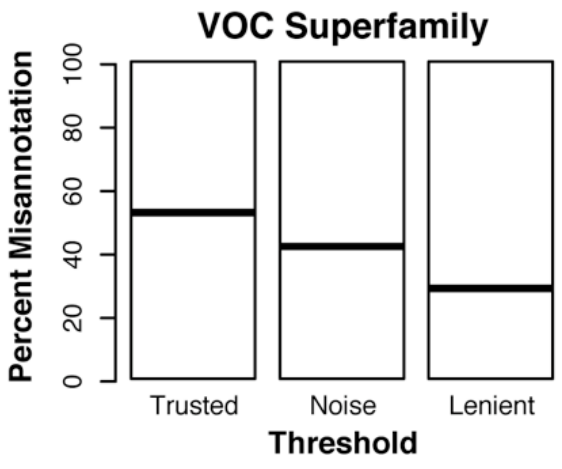
NC — Noise Cutoff

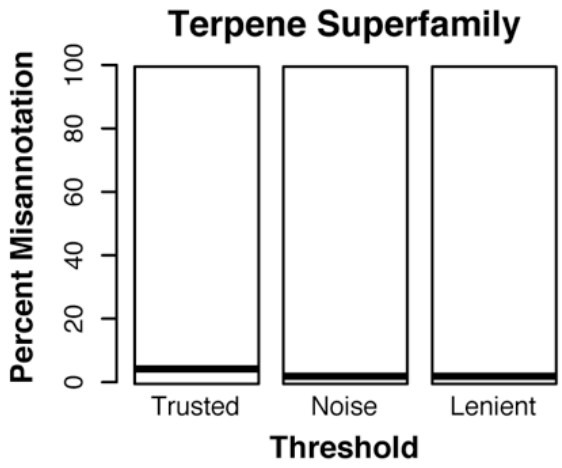
LC — Lenient Cutoff

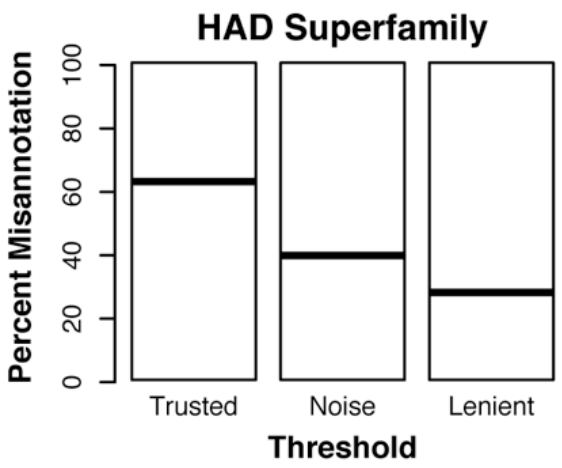
Sensitivity to threshold change

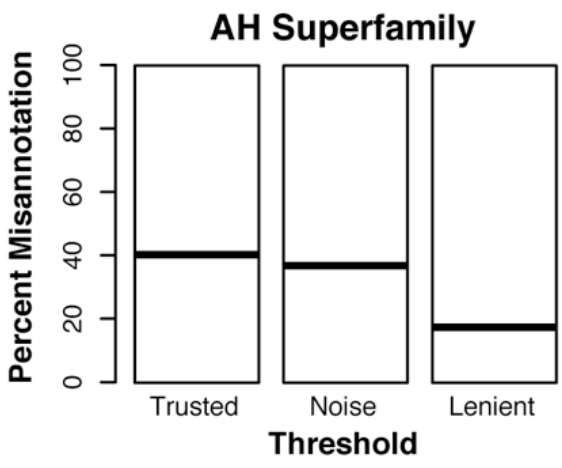


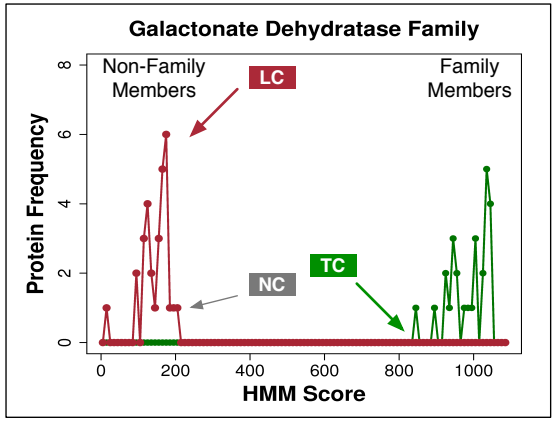




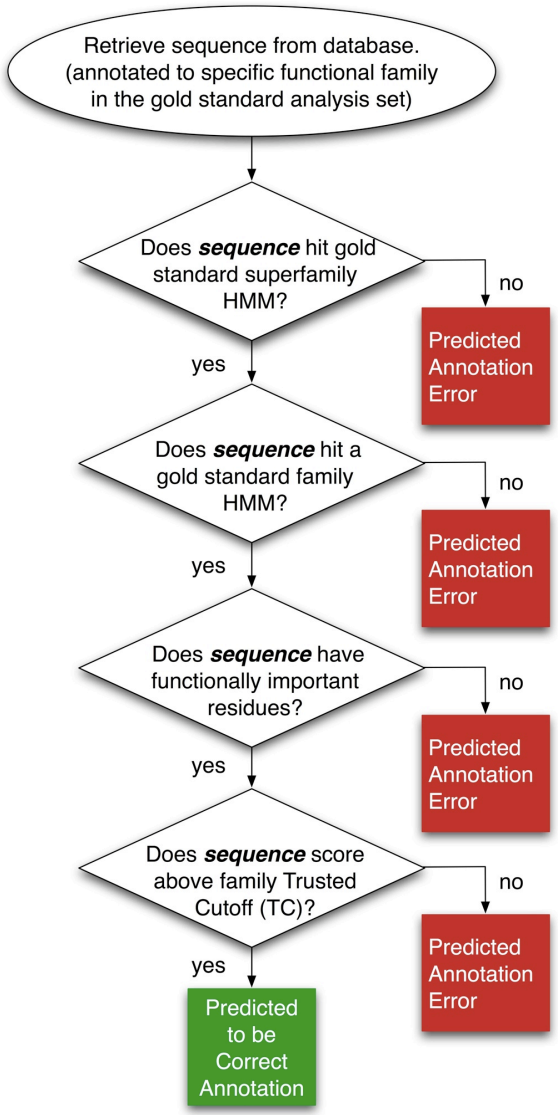




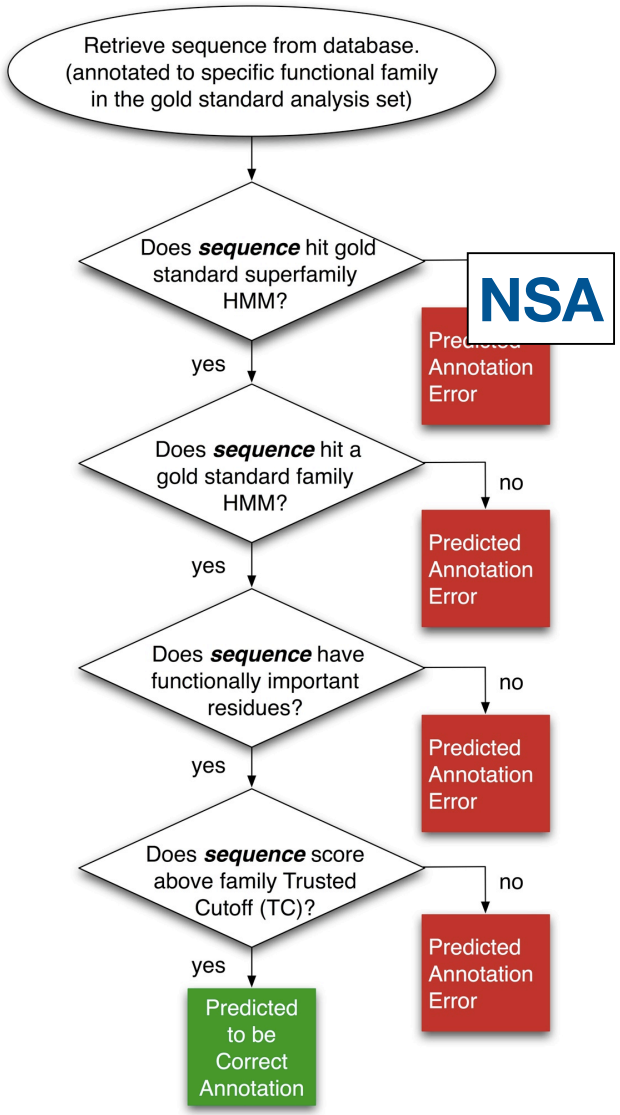




Sensitivity to threshold change

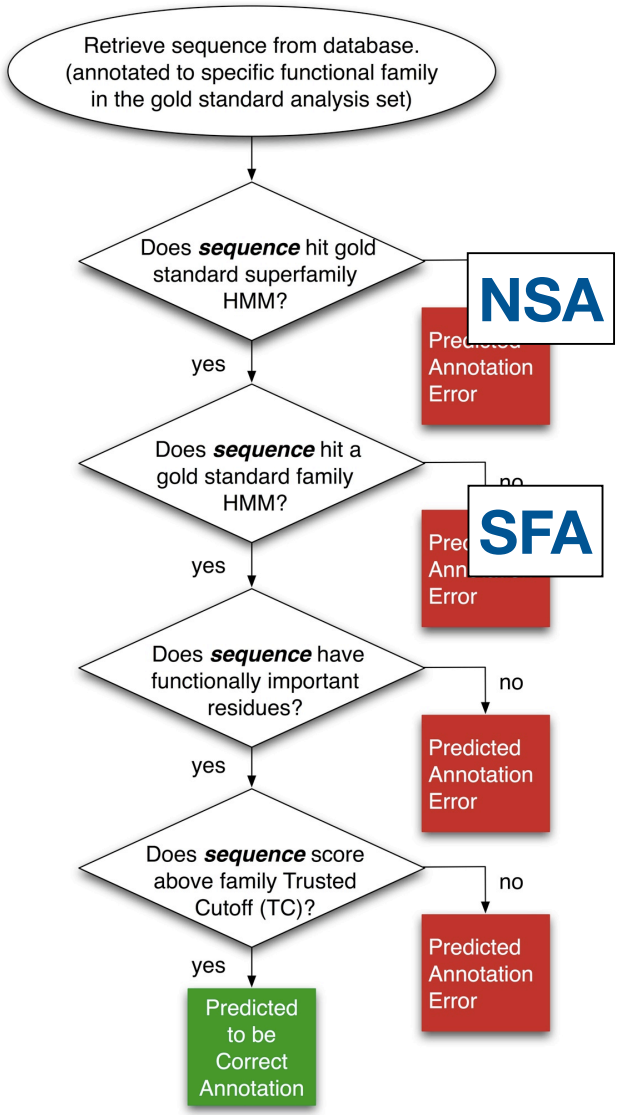

Retrieve sequence from database. (annotated to specific functional family in the gold standard analysis set)
Does sequence hit gold standard superfamily HMM?
no
Predicted Annotation Error
yes
Does sequence hit a gold standard family HMM?
no
Predicted Annotation Error
yes
Does sequence have functionally important residues?
no
Predicted Annotation Error
yes
Does sequence score above family Trusted Cutoff (TC)?
no
Predicted Annotation Error
yes
Predicted to be Correct Annotation

NSA — No Superfamily Association



NSA — No Superfamily Association

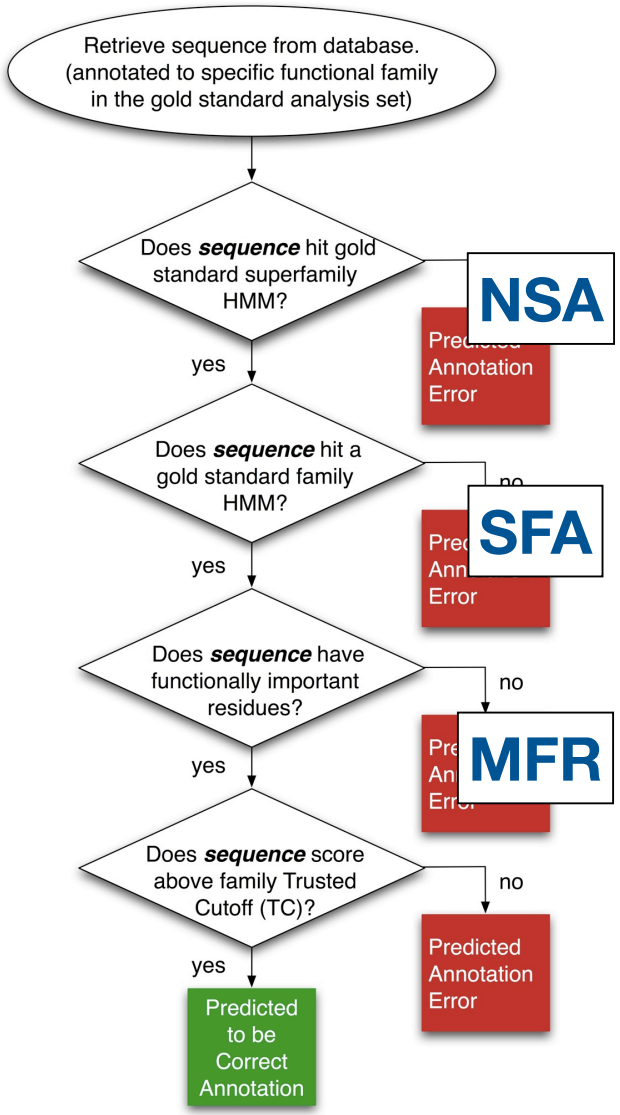
SFA — Superfamily Association Only



NSA — No Superfamily Association

SFA — Superfamily Association Only

MFR — Missing Functionally Important Residues

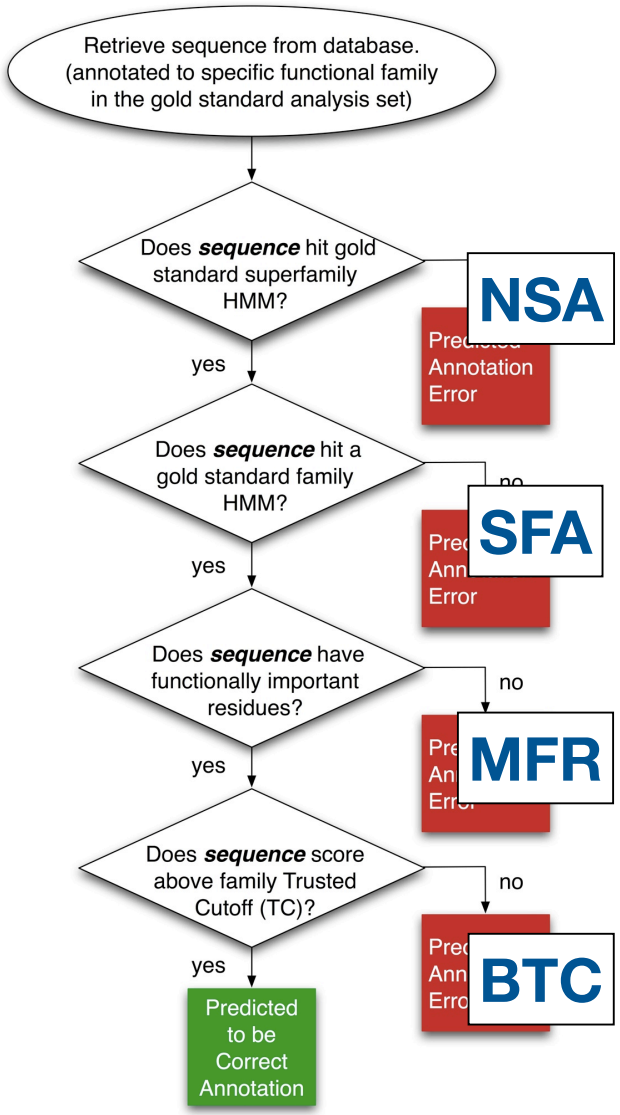


NSA — No Superfamily Association

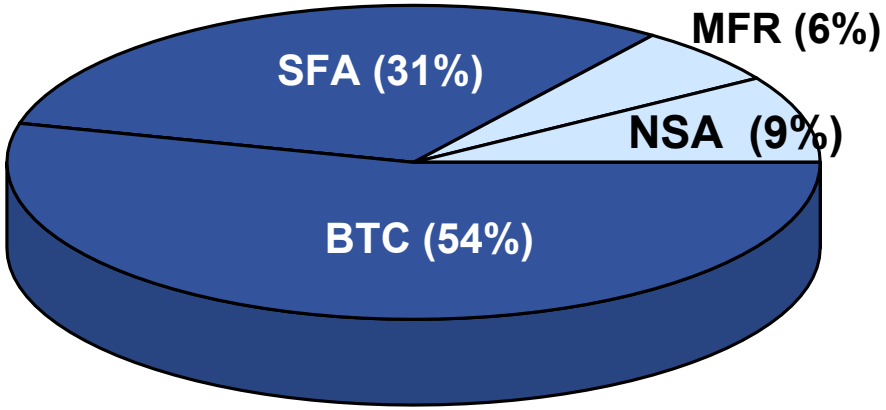
SFA — Superfamily Association Only

MFR — Missing Functionally Important Residues

BTC — Below Trusted Cutoff



Types of Misannotation



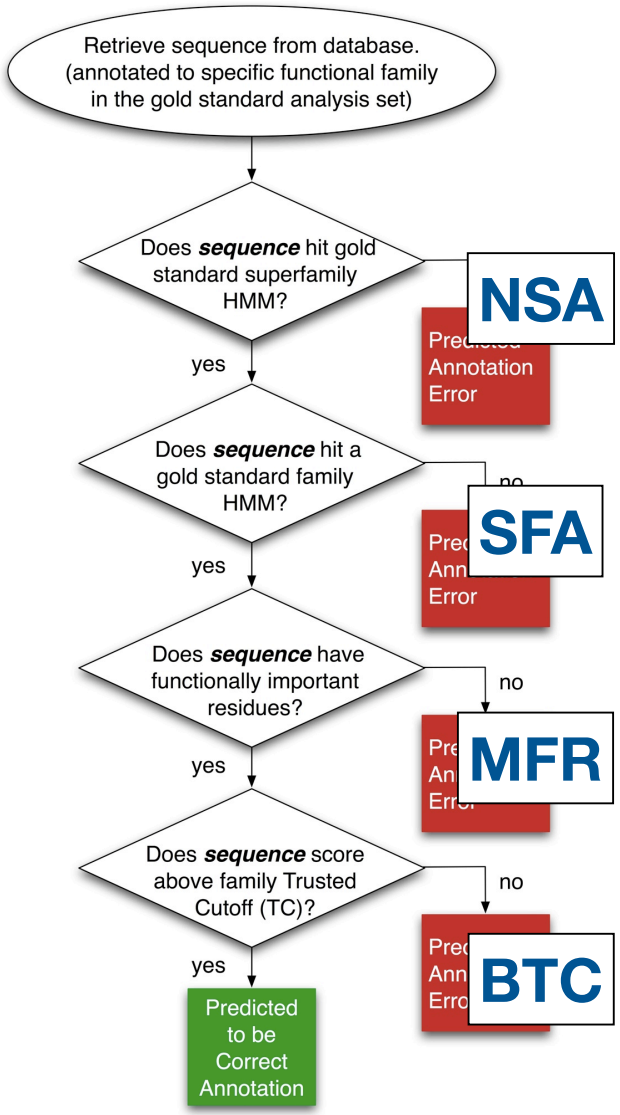
Misannotations due to overprediction
Misannotations not due to overprediction

NSA — No Superfamily Association

SFA — Superfamily Association Only

MFR — Missing Functionally Important Residues

BTC — Below Trusted Cutoff



Biggest Problem

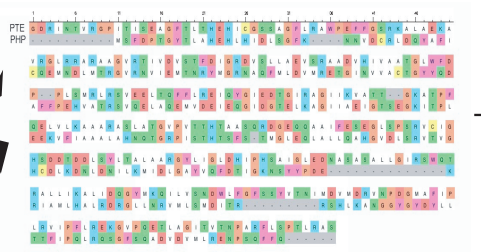
Predicting function without sufficient evidence

Dipeptide Epimerase

>gi|13786715|pdb|1HZY|A Ch… Resolution Structure Of The Zinc-Containing Phosphotriesterase From Pseudomonas Diminuta
GDRINTVRGPITISEAGFTL… EHIC… GFLRA…PEFFGSRKALAEKAVRGLRRARAAGVRTIVDVST
FDIGRDVSLLAEVSRAADL… HIVAAT… FDPPLS… RLRSVEELTQFFLREIQYGIEDTGIRAGIIKVATT
GKATPFQELVLKAAARAS… ATGVPVT… TAASQF… GEQQAAIFESEGLSPSRVCIGHSDDTDDLSYLTAL
AARGYLIGLDHIPHSAIGL… DNASAS… GIRSW… RALLIKALIDQGYMKQILVSNDWLFGFSSYVTNI
MDVMDRVNPDGMAFIPL… IPFLRE… VPQET… AGITVTNPARFLSPTLRAS

>gi|1176259|sp|P45548|PL… ECOLI Phosphotriesterase homology protein
MSFDPTGYTLAHEHLHID… GFL… RLDQ… FICQEMNDLMTRGVRNVIEMTNRYMGRNAQFMLDVM
RETGINVVACTGYYQDA… PEHVATRS… ELAQEL… VDEIEQGIDGTELKAGIIAEIGTSEGKITPLEEKV
FIAAALAHNQTGRPIST… SFSTMG… LALLQAH… GVDLSRVTVGHCDLKDNLDNILKMIDLGAYVQFDT
IGKNSYYPDEKRIAMLHA… RDRG… VMLSMDI… RRSHLKANGGYGYDYLLTTFIPQLRQSGFSQADVD
VMLRENPSQFFQ

Unknown Function



Dipeptide Epimerase

① = ②

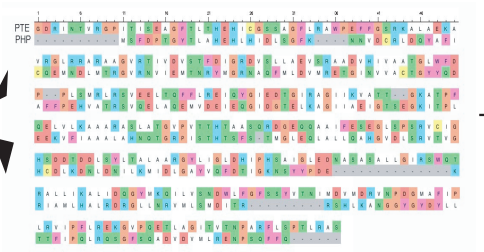
Dipeptide Epimerase

Error Propagation

Dipeptide Epimerase

>gi|13786715|pdb|1HZY|A Ch… Resolution Structure Of The Zinc-Containing Phosphotriesterase From Pseudomonas Diminuta
GDRINTVRGPITISEAGFTL… EHIC… GFLRA… PEFFGSRKALAEKAVRGLRRARAAGVRTIVDVST
FDIGRDVSLLAEVSRAADL… HIVAAT… FDPPLS… RLRSVEELTQFFLREIQYGIEDTGIRAGIIKVATT
GKATPFQELVLKAAARAS… ATGVPVT… TAASQF… GEQQAAIFESEGLSPSRVCIGHSDDTDDLSYLTAL
AARGYLIGLDHIPHSAIGL… DNASAS… GIRSW… RALLIKALIDQGYMKQILVSNDWLFGFSSYVTNI
MDVMDRVNPDGMAFIPL… IPFLRE… VPQET… AGITVTNPARFLSPTLRAS

>gi|1176259|sp|P45548|PL… ECOLI Phosphotriesterase homology protein
MSFDPTGYTLAHEHLHID… GFL… RLDQ… FICQEMNDLMTRGVRNVIEMTNRYMGRNAQFMLDVM
RETGINVVACTGYYQDA… PEHVATRS… ELAQEL… VDEIEQGIDGTELKAGIIAEIGTSEGKITPLEEKV
FIAAALAHNQTGRPISTH… SFSTMG… LALLQAH… GVDLSRVTVGHCDLKDNLDNILKMIDLGAYVQFDT
IGKNSYYPDEKRIAMLHA… RDRG… VMLSMDI… RRSHLKANGGYGYDYLLTTFIPQLRQSGFSQADVD
VMLRENPSQFFQ



Dipeptide Epimerase

① = ②

Dipeptide Epimerase

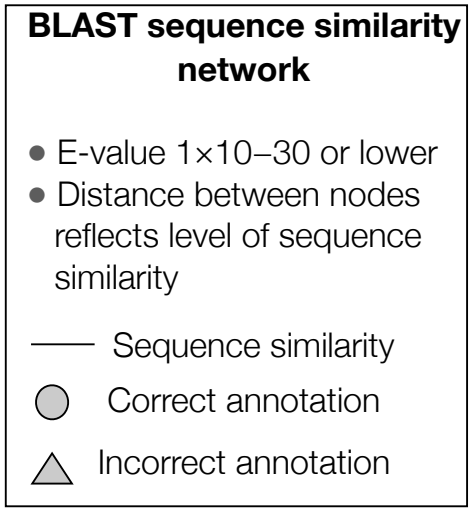
Unknown Function

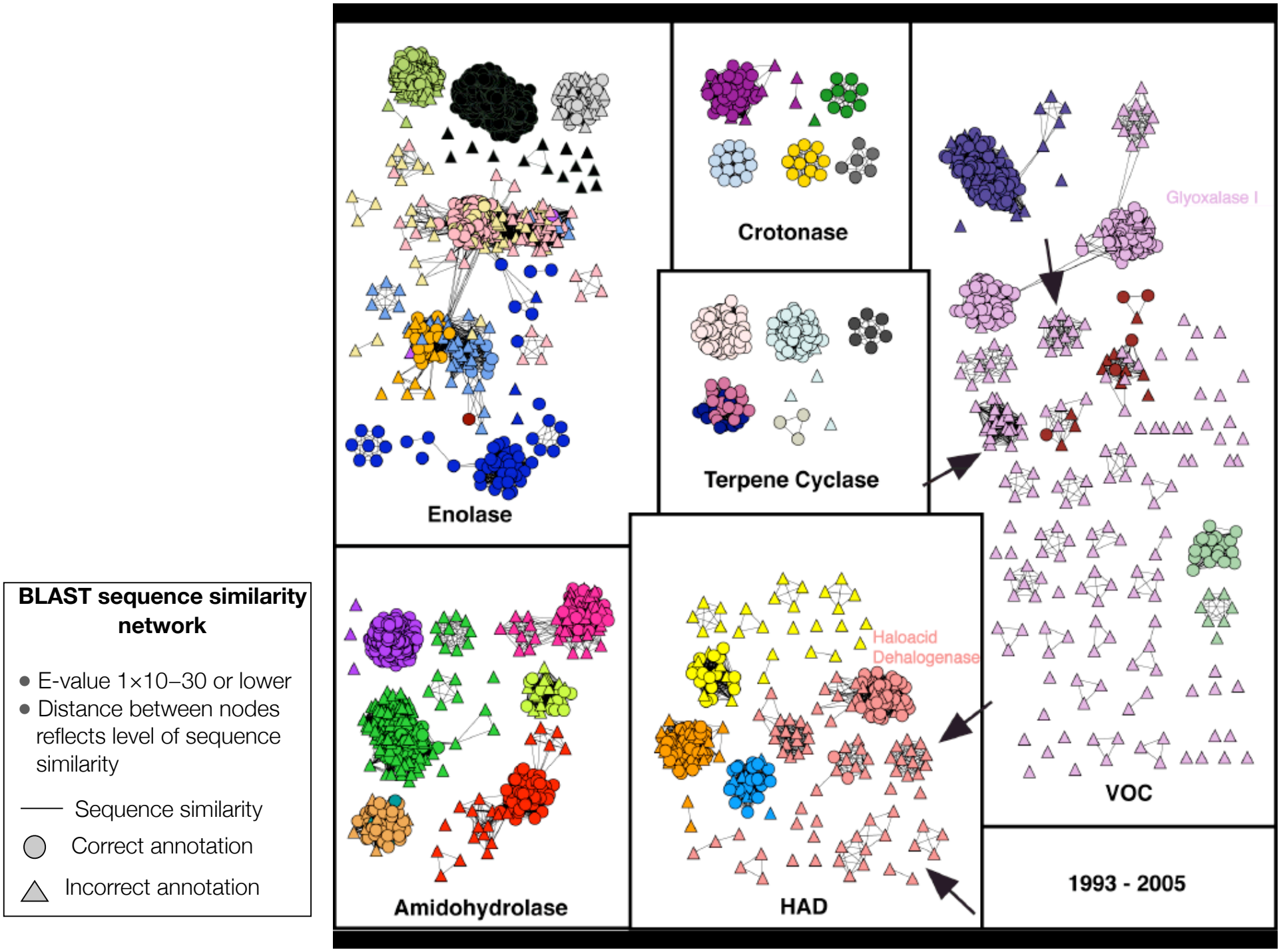
① ≠ Dipeptide Epimerase

① & ②

INCORRECT!

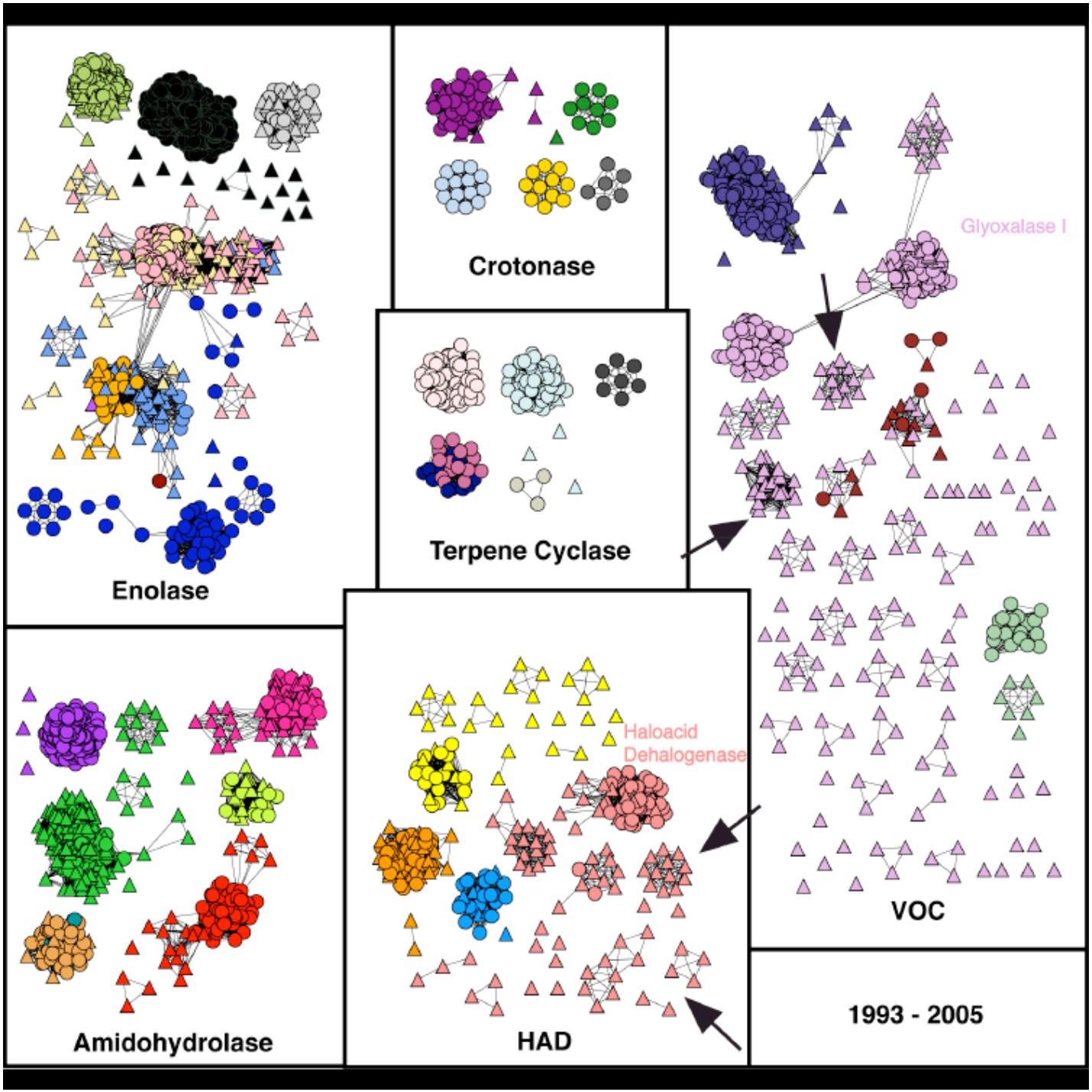
Error Propagation


BLAST sequence similarity network
E-value 1×10–30 or lower
Distance between nodes reflects level of sequence similarity
Sequence similarity
Correct annotation
Incorrect annotation


BLAST sequence similarity network
E-value 1×10−30 or lower
Distance between nodes reflects level of sequence similarity
Sequence similarity
Correct annotation
Incorrect annotation
Enolase
Crotonase
Terpene Cyclase
Glyoxalase I
VOC
Amidohydrolase
HAD
Haloacid Dehalogenase
1993 - 2005

Misannotations

- Cluster with each other
- Indication of error propagation



BLAST sequence similarity network

- E-value 1×10−30 or lower
- Distance between nodes reflects level of sequence similarity

Sequence similarity

Correct annotation

Incorrect annotation

In Conclusion...

- Misannotation is a serious problem
  - Automated databases
  - Across multiple folds, functions and superfamilies
  - Hard to predict misannotation *a priori*
  - Manual curation delivers the highest quality
- Misannotation problem is getting worse
- Overprediction is a common problem
- Error propagation appears to be a common source of misannotation

Acknowledgements

Patricia Babbitt & lab
Shoshana Brown

Igor Dodevski
University of Zürich

Tanja Kortemme & Lab
Colin Smith

Jim Wells Lab
Emily Crawford

$$ Howard Hughes Pre-Doctoral Fellowship
NIH & NSF

PLoS Comput Biol. 2009 Dec;5(12):e1000605.

Wiki Commons & Science Magazine for some images