

# Using the Superfamily Perspective for Inference of Protein (Molecular) Function

Mechanistically diverse enzyme superfamilies

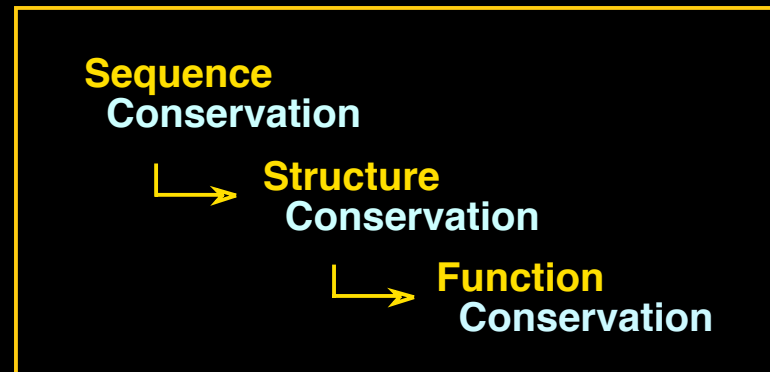
Patsy Babbitt  
UCSF  
May 12, 2004

## Why do we care?

- Function is the only element of the Sequence → Structure → Function paradigm that we don't really know how to address either theoretically or computationally
- Inference of molecular function/functional characteristics remains a primary unsolved problem for post-genome era informatics

Genome	# of Predicted ORFS	% Similar to Unknowns	% Not Similar to Anything	Total % URFs
P. horikoshii	2061	22	58	80
A. fulgidus	2436	~25	~25	50
M. jannaschii	1738	-	-	62
E. coli	4288	-	-	38
C. elegans	19099	-	-	~55
S. cerevisiae	6183	-	-	>35
D. melanogaster	13601	-	23	?
H. sapiens	~30,000	-	26	?

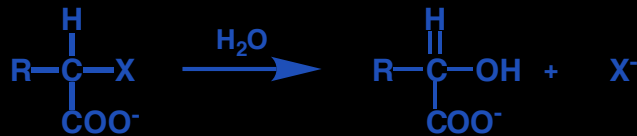
# Traditional approaches for inference of function



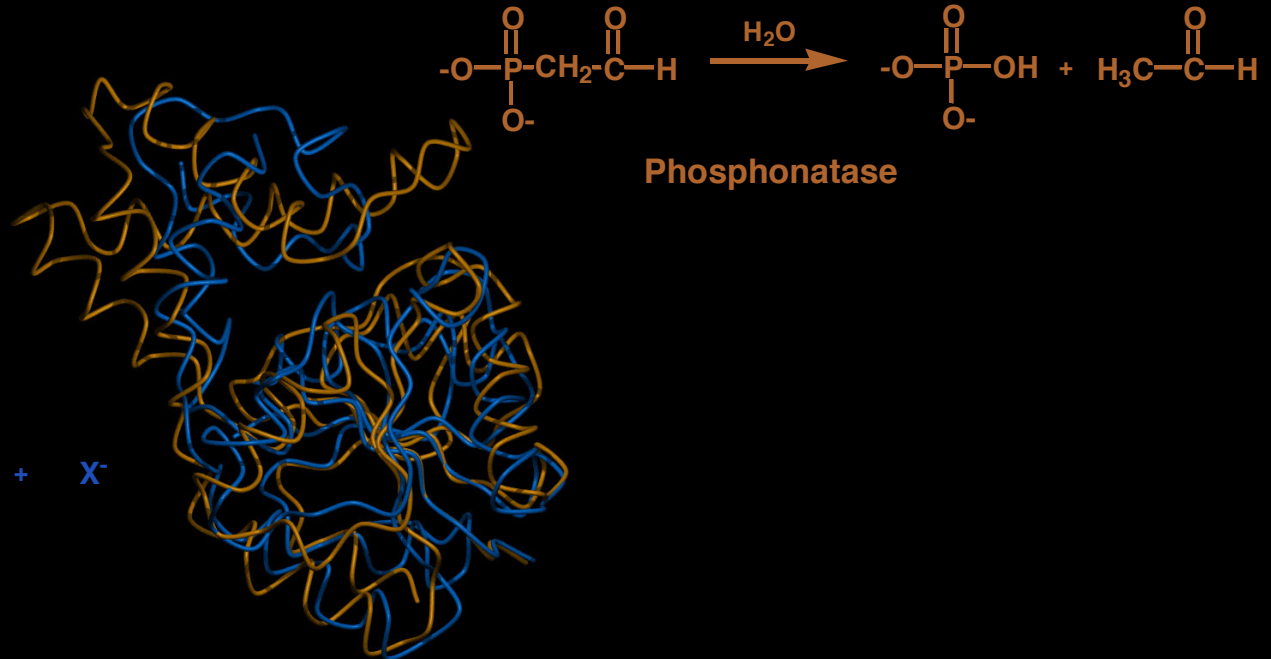
- Molecular function (but not necessarily specificity) can be inferred when homologs of sufficient similarity exist in the databases
  - Examples: serine proteases, glutathione S-transferases
- ...PROVIDING ANY DIVERGENCE OF FUNCTION HAS CHANGED SPECIFICITY BUT NOT CHEMISTRY...
- Newer, non-homology based approaches help but have their own problems

## When no statistically significant sequence relationships can be identified...

- The 3-D genome projects are invoked as the solution to functional inference
- But homologs that can only be seen/verified at the structural level have also frequently diverged to mediate very different overall functions



2-Haloalkanoic  
Dehalogenase



Phosphonatase

How do we develop rules-based inference of function from sequence/structure for these non-trivial problems?

Our approach: Look at the structural strategies nature has used to evolve new enzyme functions from a limited set of scaffolds

# Models for Functional Divergence

- What are the architectural design principles associated with delivery of function for a given structural template?
  - requires explicit mappings between conserved elements of structure and conserved elements of function
  - the level of divergence most useful to look at is the superfamily level
    - › where function has diverged sufficiently to discriminate the specific aspects of function associated with conserved elements of structure
  - are these models generalizable?
  - can we use them predictively for inference of function?

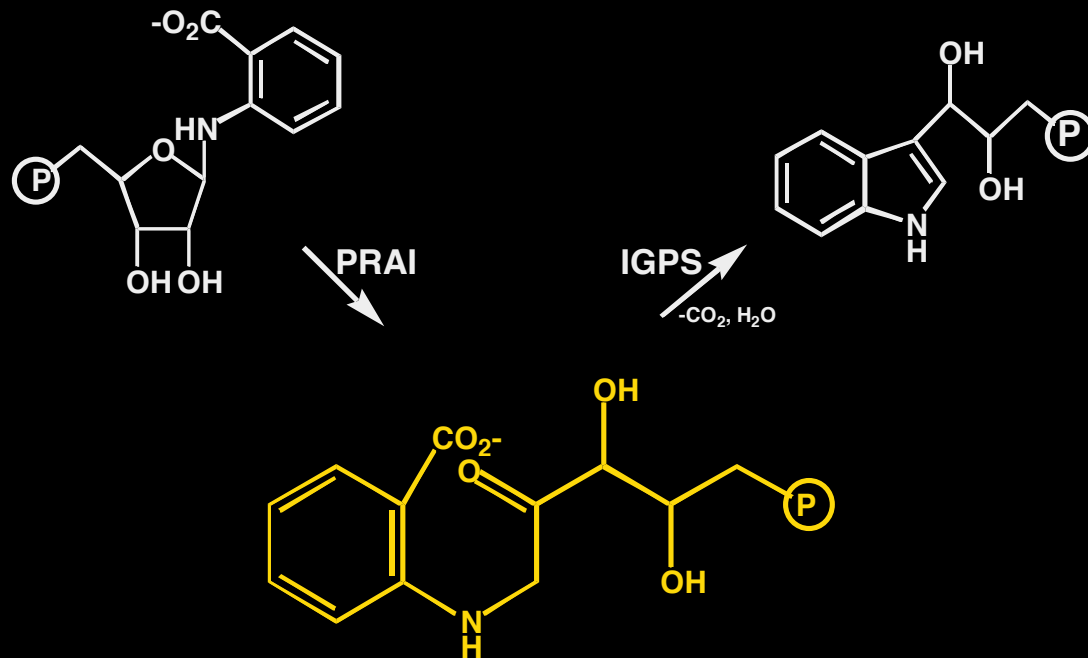
# Retro (substrate constrained) evolution

- Substrate binding determinants are dominant

Horowitz, PNAS 31 (1945) 153-157;

Horowitz, in *Evolving Genes & Proteins* (1965) 15-23

- structural elements involved in substrate binding are conserved across divergent proteins while new chemistries evolve



# Chemistry-constrained evolution

Jensen, R.A. Ann. Rev. Microbiol. 30:409-425 (1976)  
Petsko, G. A. et al., TIBS 18: 372-376 (1993)  
Babbitt & Gerlt, J. Biol. Chem. 272:30591-4 (1997)  
Gerlt & Babbitt, Ann. Rev. Biochem. 70: 209-246 (2001)

- structural elements involved in mediating chemistry are conserved across divergent proteins while the ability to bind new substrates evolves
- characteristics of chemistry-constrained evolution in mechanistically diverse enzyme superfamilies
  - › Recognizable in superfamilies of highly divergent enzymes
  - › Related proteins may differ substantially in substrates/ products and in overall function
  - ➡ Each functionally distinct member of a superfamily shares a common fundamental step in its **chemical** mechanism that can be mapped explicitly to conserved elements of structure
- substantial evidence is accumulating in support of this model
- contrasted with “substrate-constrained” evolution, in which ligand binding determinants are conserved and chemistry evolves
- in many large superfamilies, both substrate constrained and chemistry constrained models may apply



# Active site constrained evolution

- Active site structure is dominant

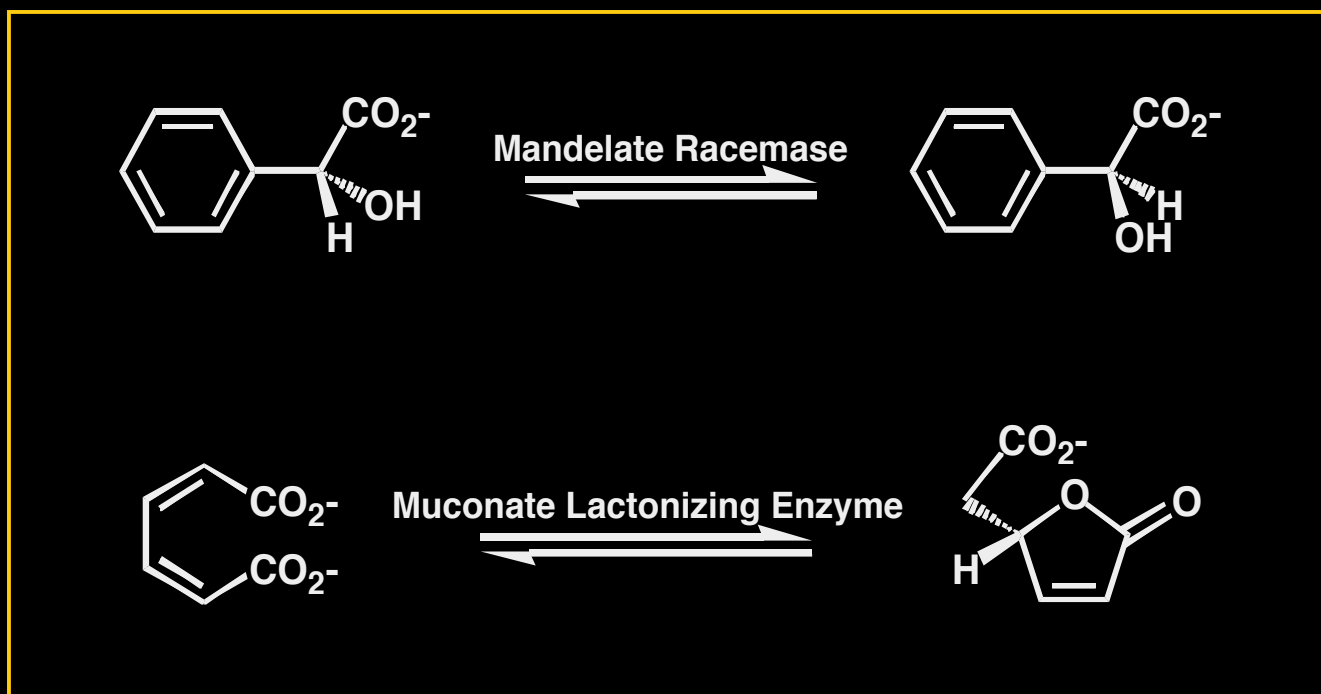
Wise, E., Biochem. 41: 3861-6 (2002)

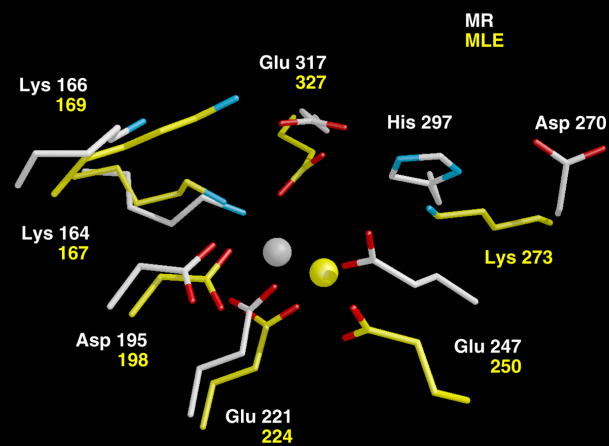
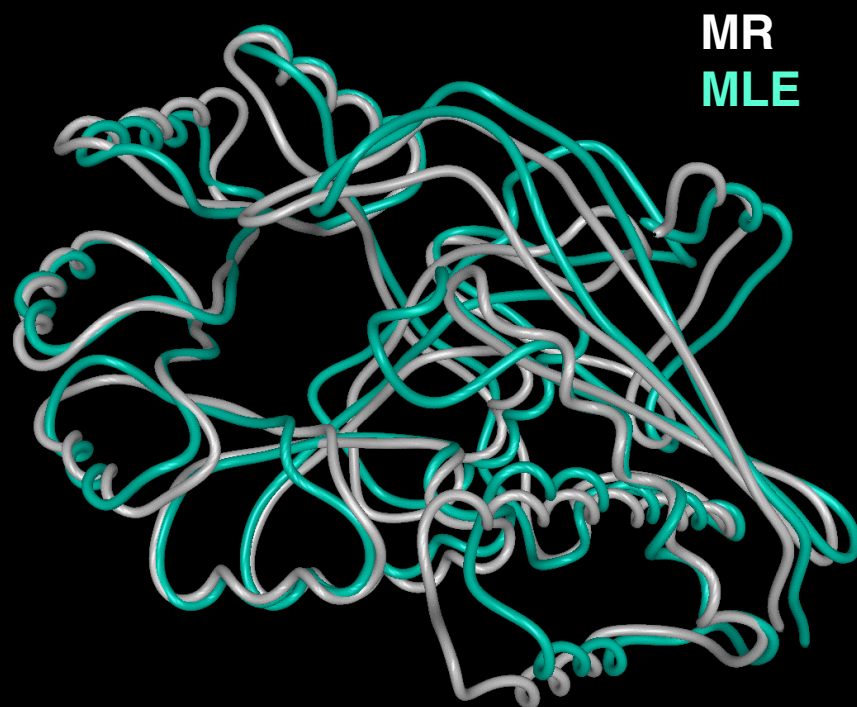
- Active site structure is conserved but used for different mechanistic steps in the overall catalytic mechanisms, e.g., no common partial reaction
  - › OMP decarboxylase - no metal, mechanism avoids formation of an unstable anion intermediate
  - › KGP decarboxylase - metal assisted stabilization of an enediolate anion intermediate

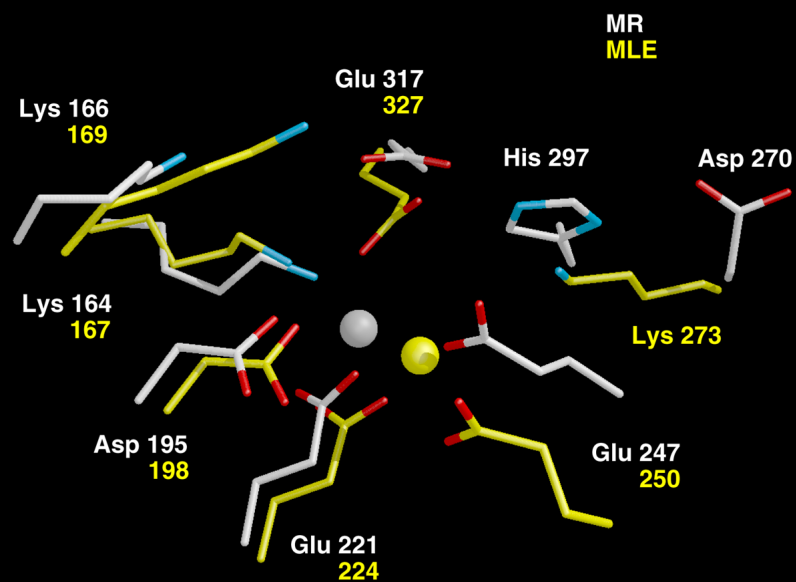


# Chemistry-constrained evolution

## Mechanistically Diverse Enzyme Superfamilies

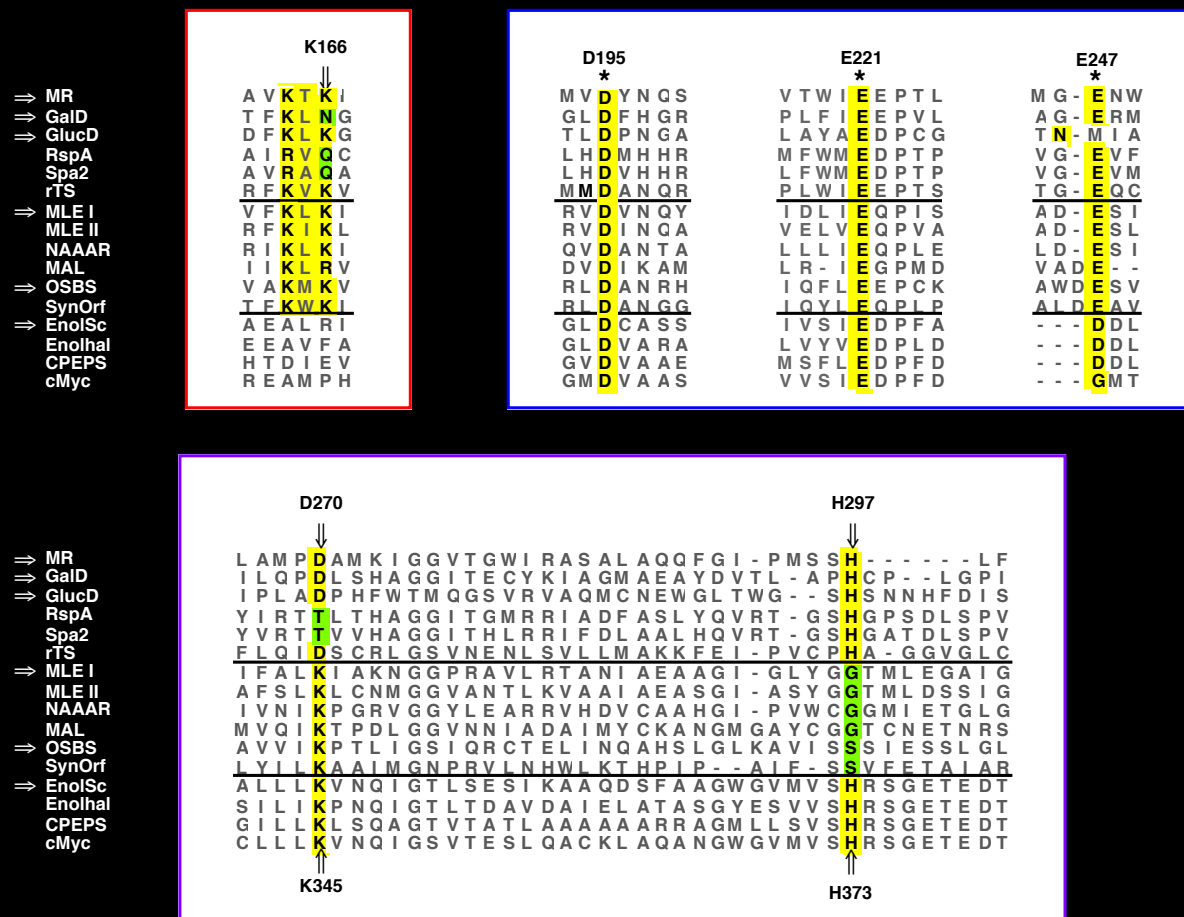


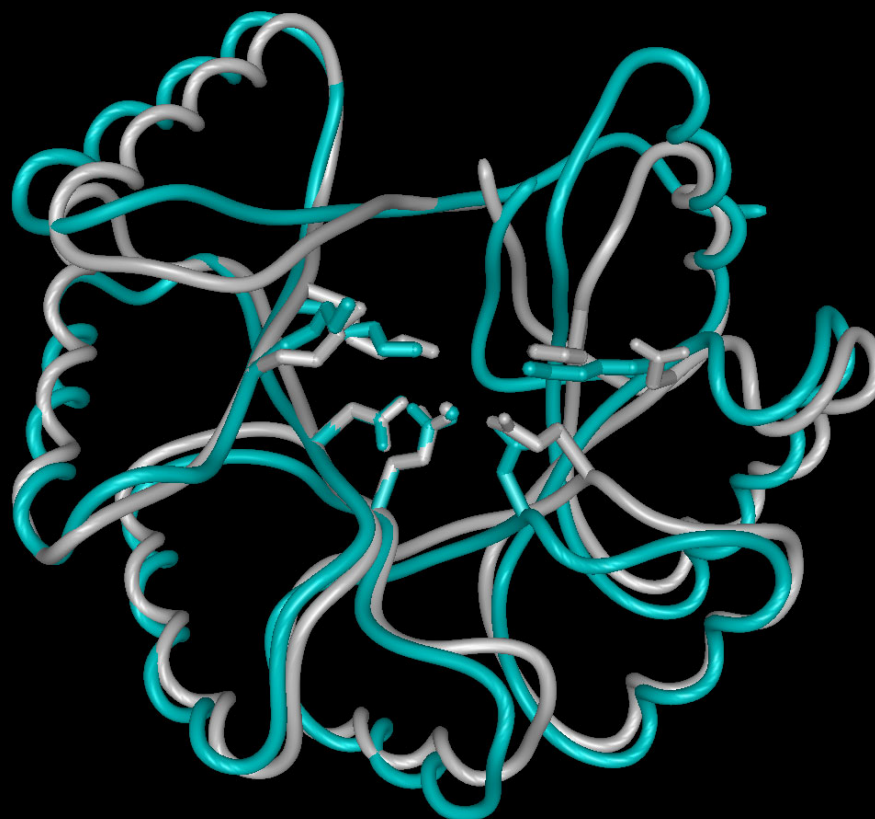
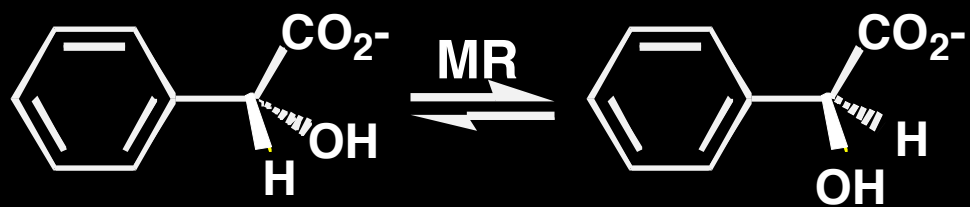


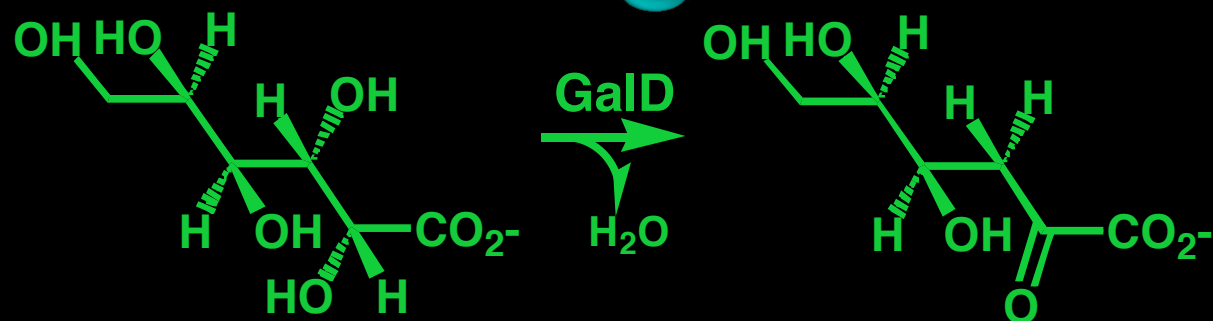
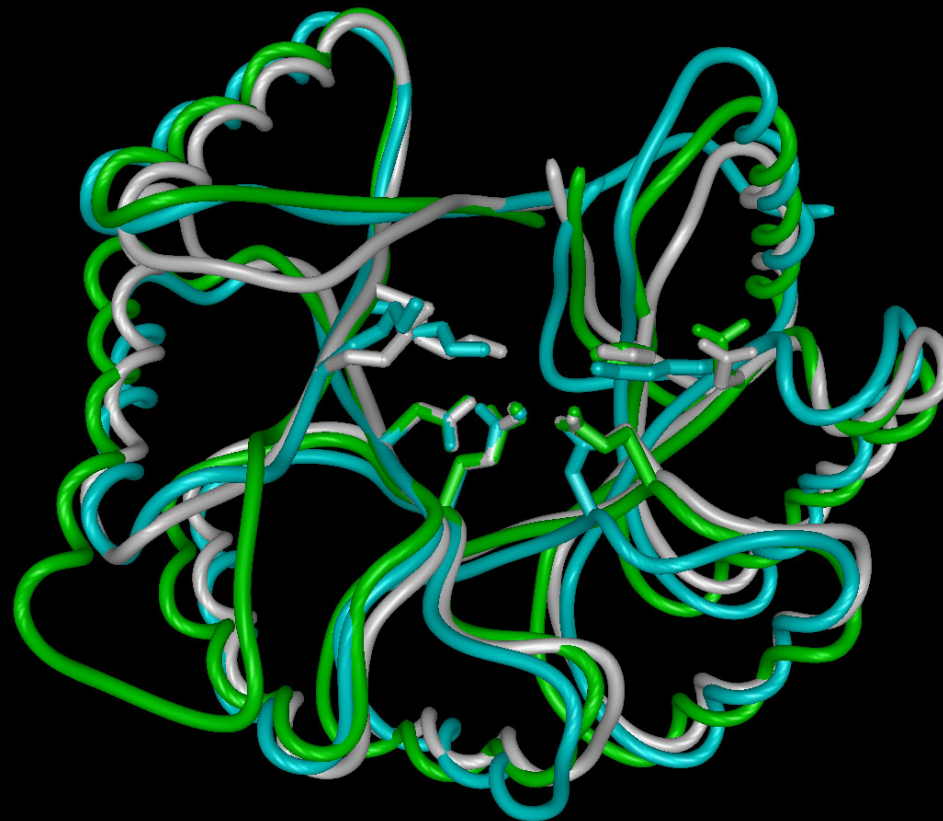


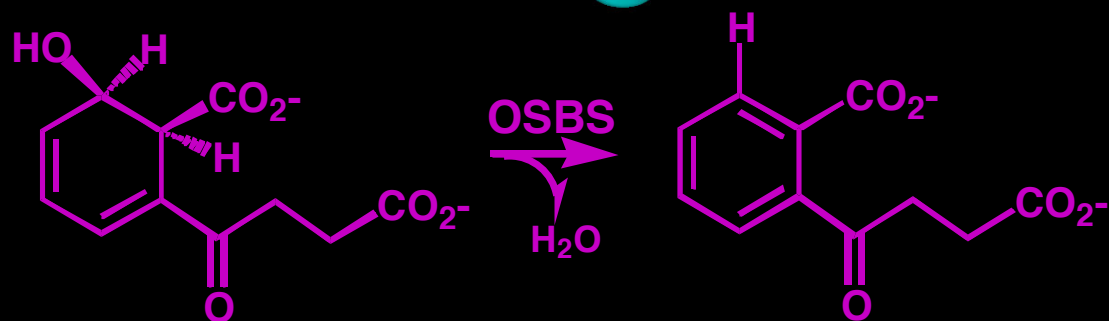
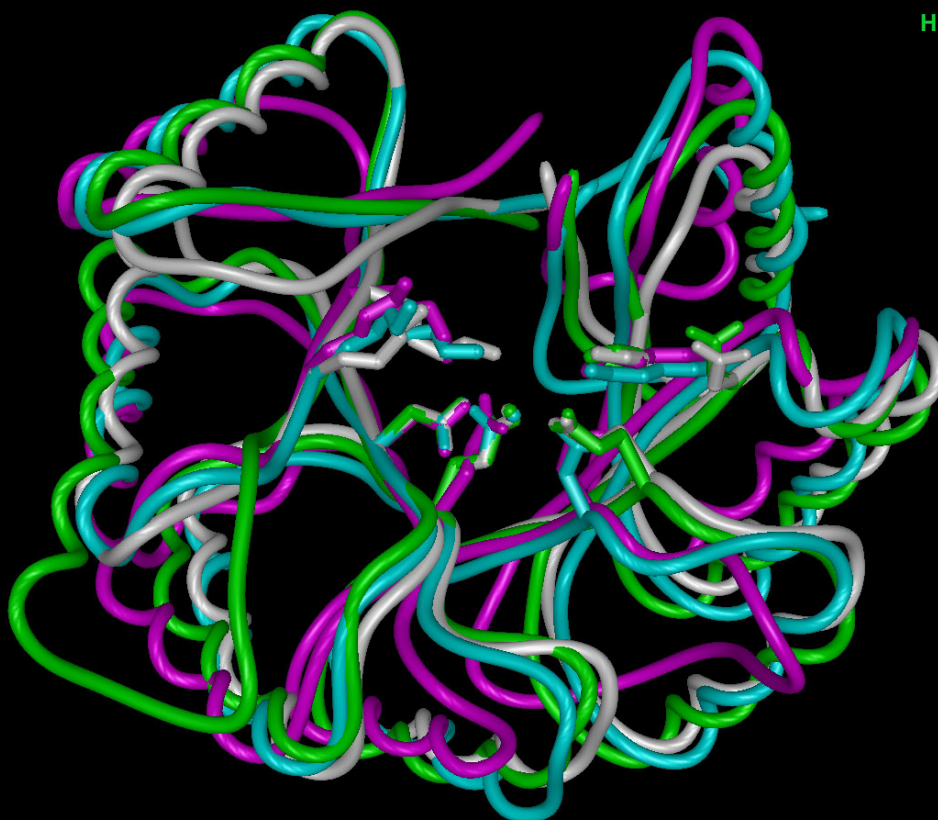
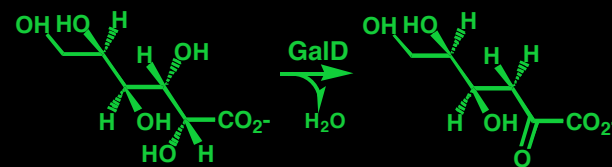
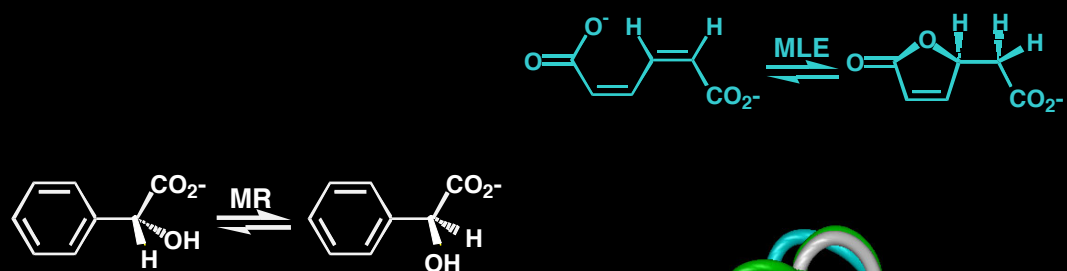
- The Enolase Superfamily now contains >500 sequences, with those representing different overall reactions showing low sequence similarity (13-35% identity in the  $\alpha/\beta$  barrel domain)
  - new “environmental” genomes from the Sargasso Sea suggest dozens of new members of unknown function
- All of these sequences share motifs representing the common active site architecture associated with the fundamental proton abstraction step

Venter et al, Science, 304 66-74 (2004)

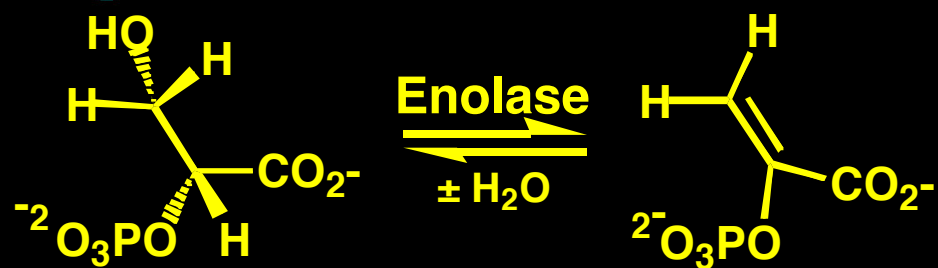
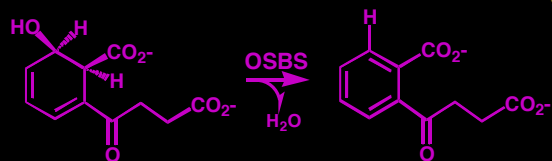
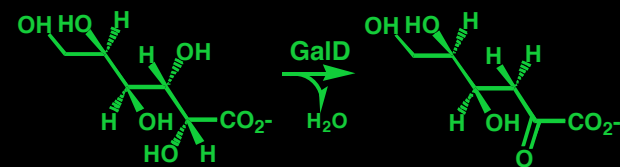
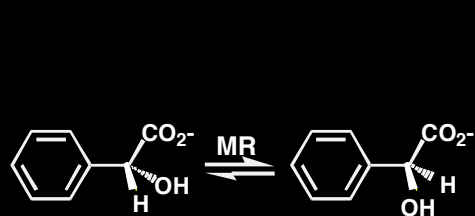


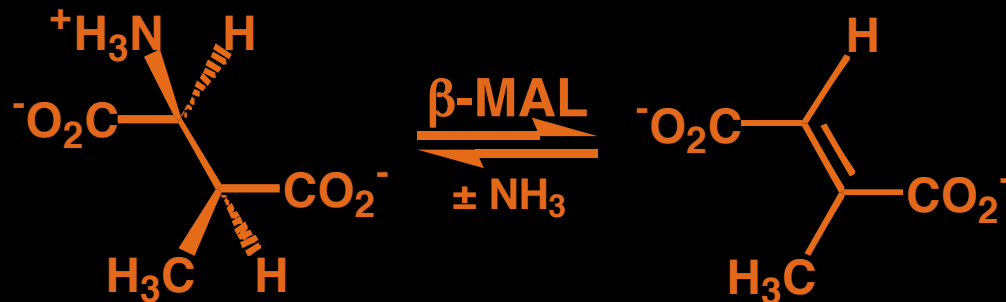
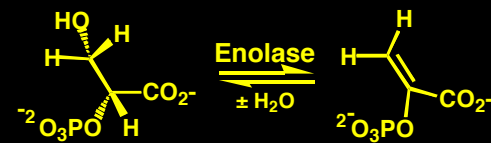
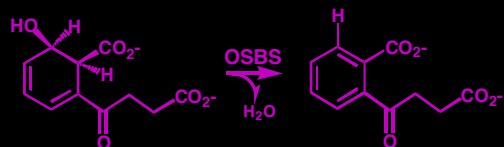
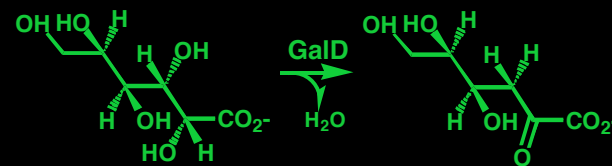
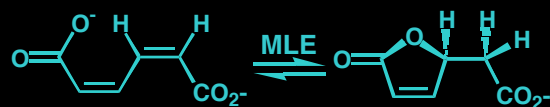


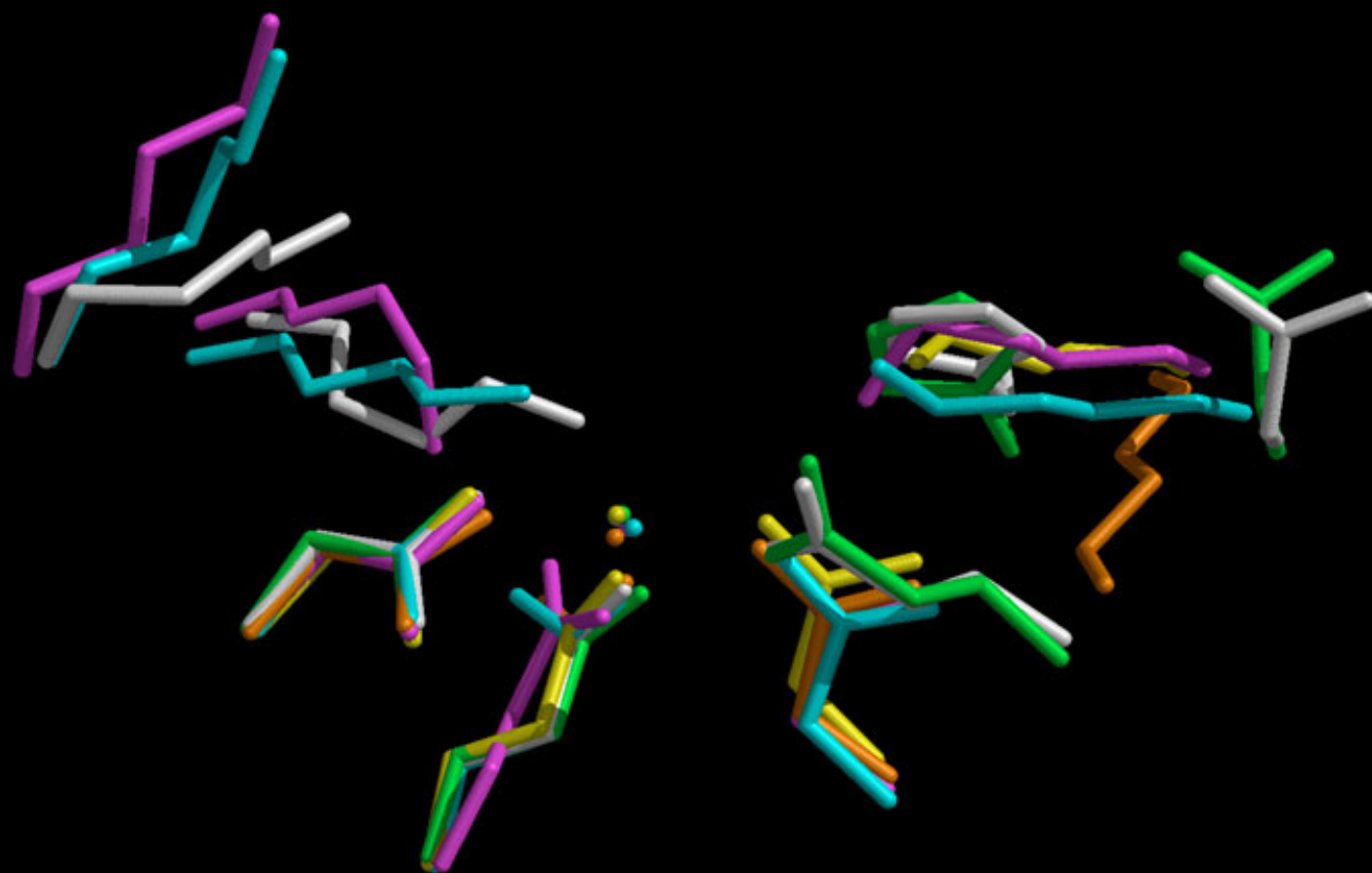




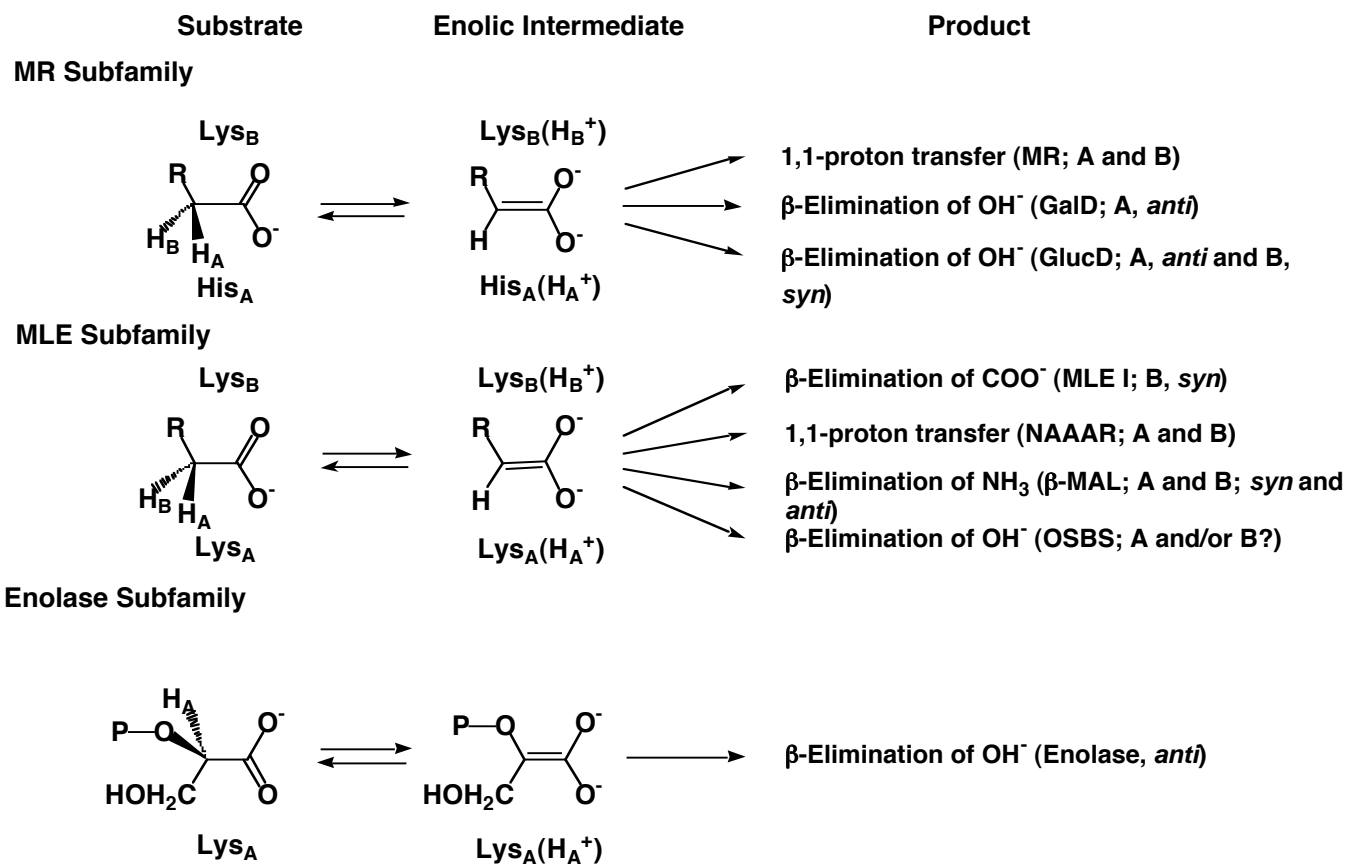
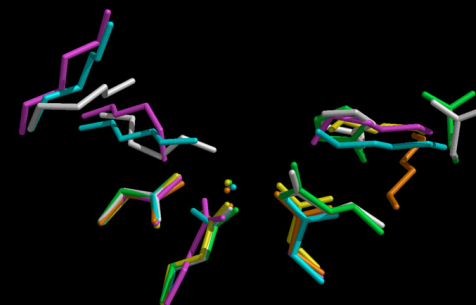




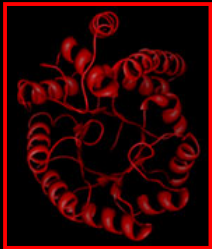




- we describe this superfamily as “a structural strategy for enzyme-catalyzed abstraction of the  $\alpha$ -protons of carboxylic acids”



# Many superfamilies fit this model



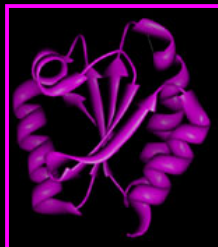
Enolase: metal dependent abstraction of  $\alpha$ -protons of carboxylic acids



Vicinal oxygen chelate: stabilization of diverse oxyanion intermediates



Crotonase: stabilization of oxyanion intermediates derived from thioesters



Haloacid dehalogenase: hydrolytic nucleophilic substitution

•  
•  
•

## Using the superfamily paradigm

- Inference of function
  - Galactonate dehydratase (enolase superfamily)
  - 4-Chlorobenzoate dehalogenase (crotonase superfamily)
- Extension of function
  - glucarate dehydratase (enolase superfamily)
- Prediction of active site residues of other ORFs/ URFs
  - RspA, Spa2, rtsA, rtsB, CPEPS (enolase superfamily)
  - a new class of diol dioxygenases (VOC superfamily)
  - MosB protein (NAL superfamily)
- Understanding chemical mechanism
  - Phosphonatases (haloacid dehalogenase superfamily)
  - Creatine kinase (guanidino kinase superfamily)
- Finding homologs using active site templates
- Correction of function
  - 2,6-Dichlorohydroquinone dioxygenase (VOC) superfamily
  - database annotation
- Protein Engineering *in vitro*
  - MLE → OSBS ← AE Epimerase

# Inference of Function

- What is the function of Orf587?

## S-substrate:

⇒ Orf587	QIGFDTF <b>K</b> L <b>N</b> GCEEL
MLE	IRRHRVF <b>K</b> L <b>K</b> IGADP
MR	ELGFRAV <b>K</b> T <b>K</b> IGYPA

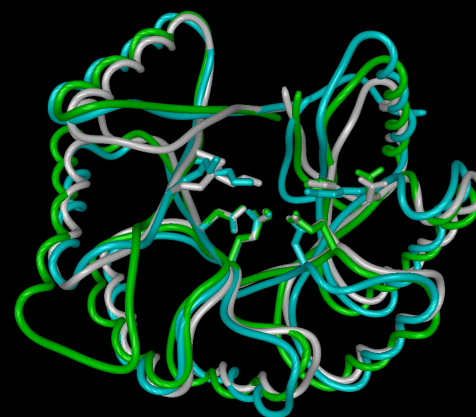
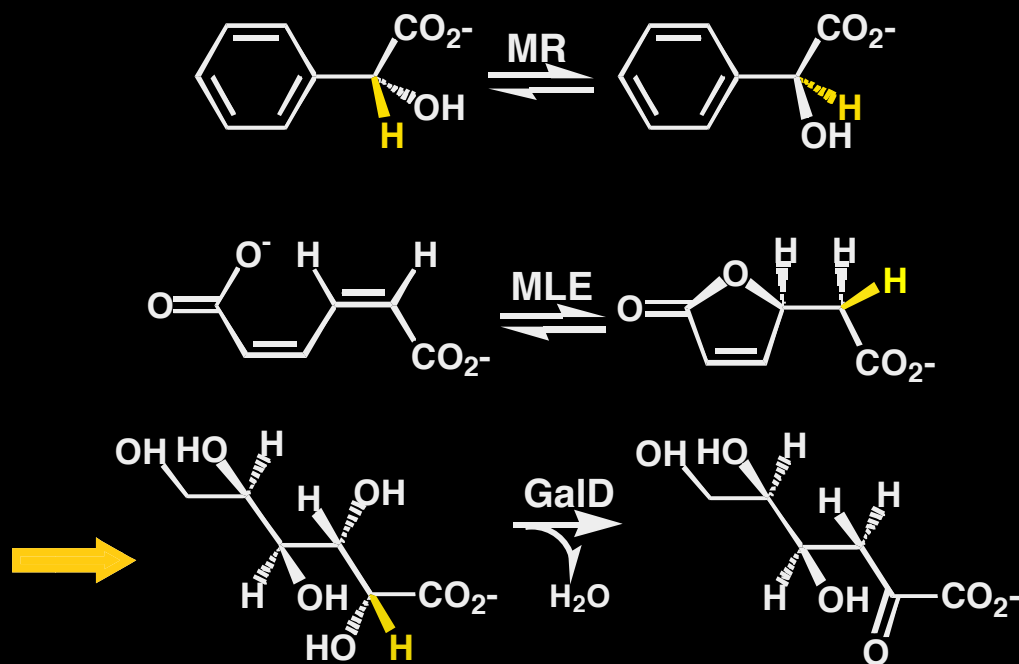
## Metal Binding:

⇒ Orf587	L <b>D</b> FHGRVSAPMAKVLIKELEPYRPLFI <b>E</b> EPVLAEQ-AEYYPKLAAQTH---IPLAAG <b>E</b> RM
MLE	V <b>D</b> VNQYWDESQAIRACQVLGDNGIDLI <b>E</b> QPISRIN-RGGQVRLNQDSP---APIMAD <b>E</b> SI
MR	V <b>D</b> YNQSLDVPAAIKRSQALQQEGVTWI <b>E</b> EPTLQHD-YEGHQRIQSKLN---VPVQMG <b>E</b> NW

## R-substrate:

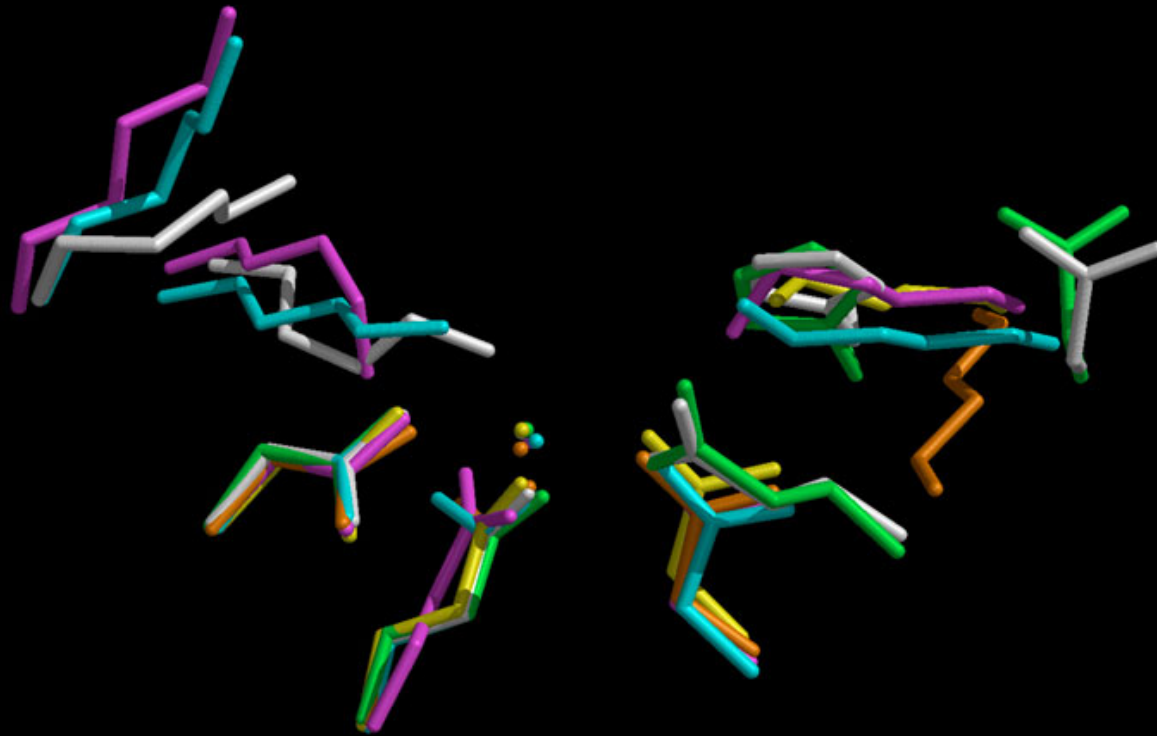
⇒ Orf587	QP <b>D</b> L-----SHAGGITECYKIAGMAEAYDVT LAP <b>H</b> CP--LGPIALAACLHIDFVSYN AVLQ <b>E</b> QS
MLE	AL <b>K</b> I-----AKNGGPRAVLRTANIAEAAGIGLYG <b>G</b> TMLEGAIGTLASAHAF LTLRQLTWGT <b>E</b> LF
MR	MP <b>D</b> A-----MKIGGV TGWIRASALAQQFGIPMSS <b>H</b> -----LFQEISAHLLAATP-TAHWL <b>E</b> RL

- Superfamily model: proton abstraction
- Subgroup clustering: stereospecificity of rxn
- Metabolic context: substrate specificity

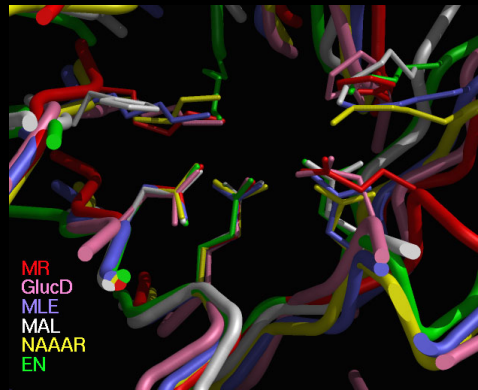




To what extent can we think of function as “hard-wired”  
into these superfamily structures?



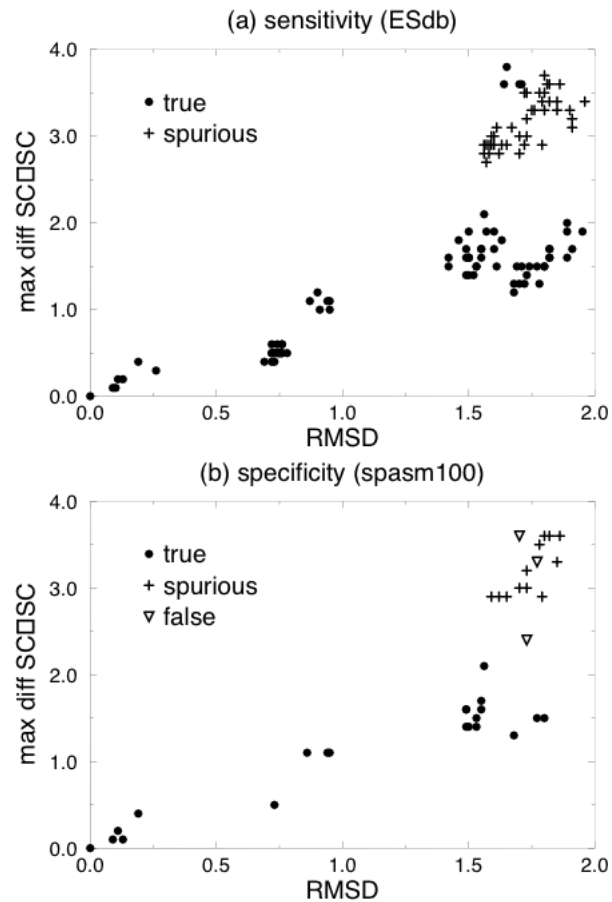
# Active site template searching for identification of new homologs and classification of subgroups/families within a superfamily



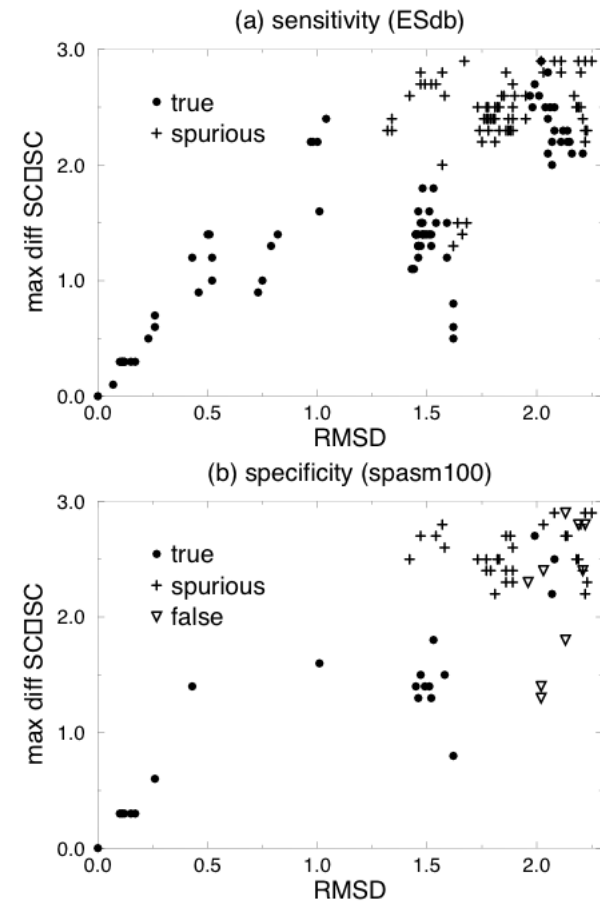
- SPASM (Spatial Arrangements of Sidechain and Mainchain)  
Kleywegt, GJ. Jour. Mol. Biol. 285: 1887-97 (1999)
  - Templates use one CA or two points per residue (CA + sidechain center of mass position)
- Databases used in the search
  - All enolase superfamily structures (44 pdb files, 83 chains) + 18 decoys closest to enolase superfamily structures
  - $(\beta/\alpha)_8$  barrels (115 non-redundant barrel domains)
  - PDB (~9,000 structures)

# Finding superfamily members

## 2mnr enolase superfamily template



## 1muc enolase superfamily template



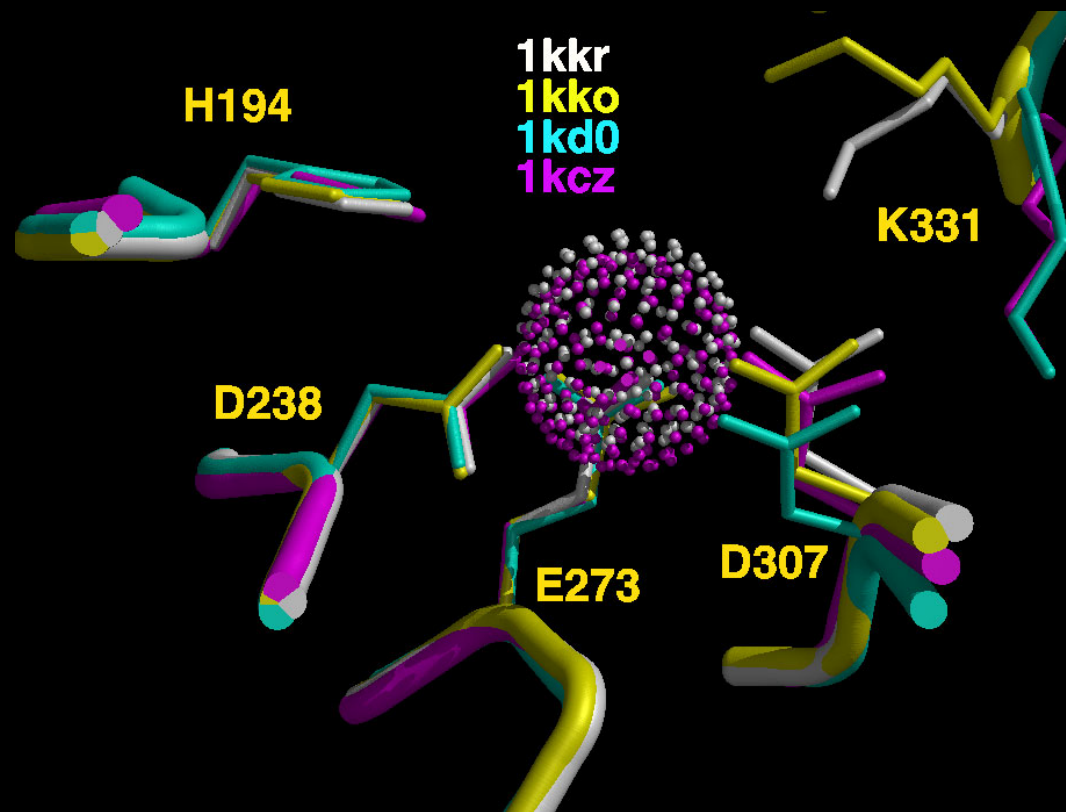
- Superfamily searches are specific and sensitive

Template	Cutoffs:			True hits, Total matches	False hits, Total matches
	RMS	CA	SC		
MR (2mnr)	2.00	5.0	3.8	44, 83 (all)	3, 3
	2.00	5.0	2.2	42, 79	0, 0
GlucD (1ec7)	2.00	5.5	3.8	44, 83	5, 5
	1.90	5.5	2.3	42, 78	0, 0
MLE (1muc)	2.25	6.5	2.9	44, 83	9, 9
AEE (1jpm)	2.20	6.1	2.8	44, 83	10, 13
MAL (1kko)	2.30	6.2	3.0	44, 83	18, 22
OSBS (1fhu)	2.20	5.1	3.6	44, 83	22, 23
Enolase (1ebh)	2.50	6.5	4.7	44, 83	95, 118
	2.30	6.5	2.7	42, 79	5, 6

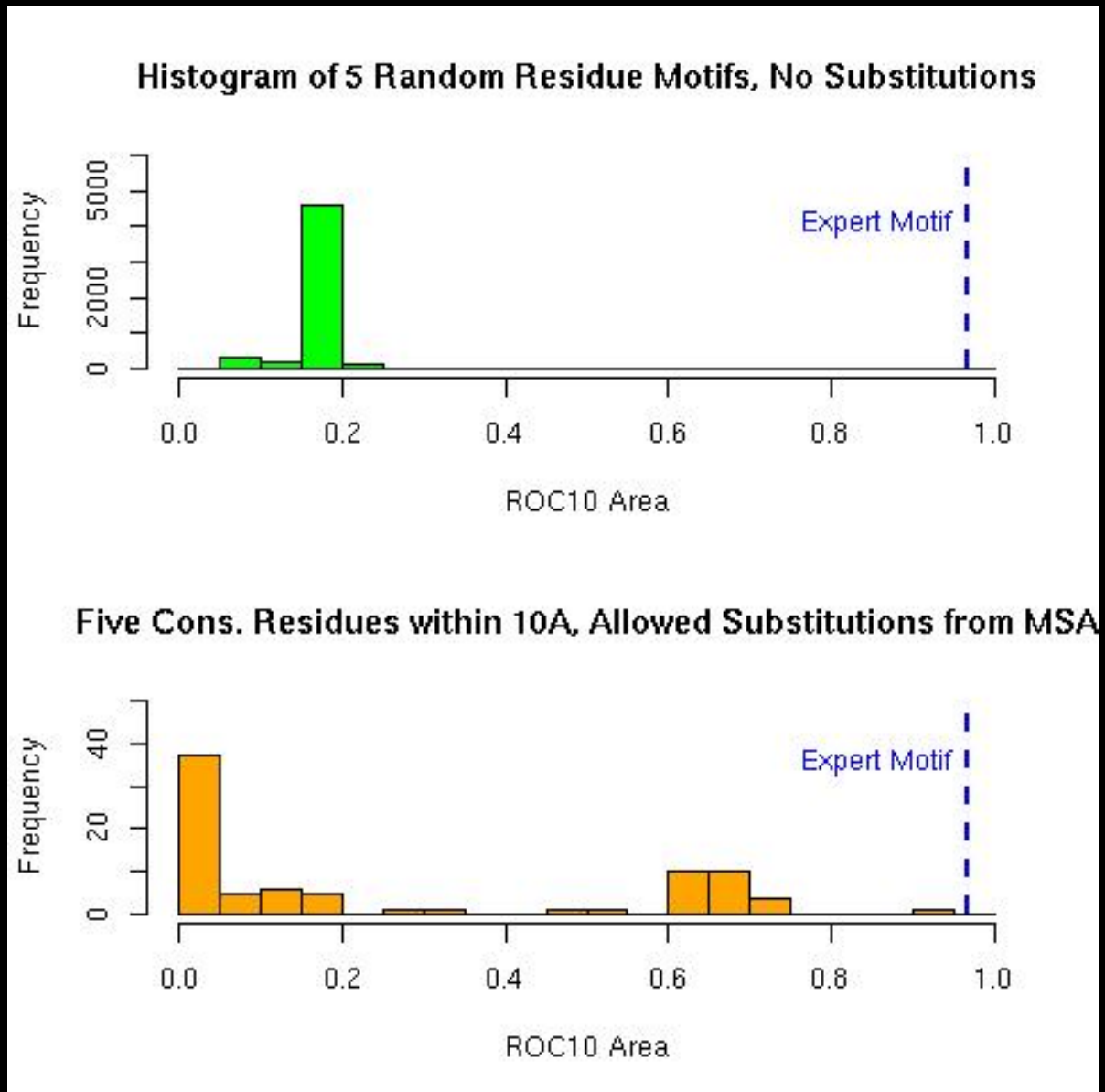
- Consensus templates do not perform as well
- Active site templates perform as well or better than FSSP, CE at finding all other superfamily members

# Where we had problems:

## Comparison of sites from MAL structures



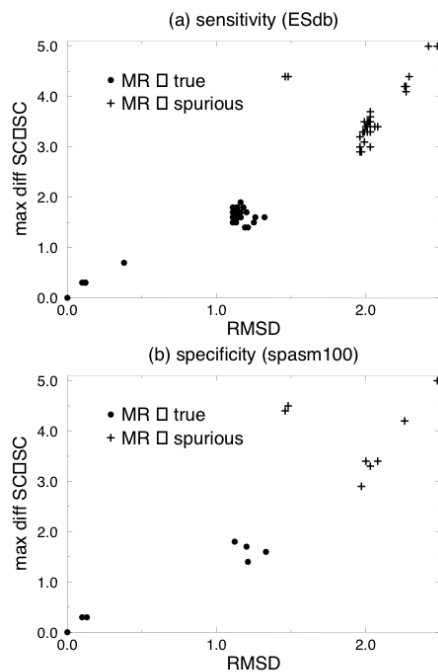
How good are our motifs compared to any 5 residues in a template structure?



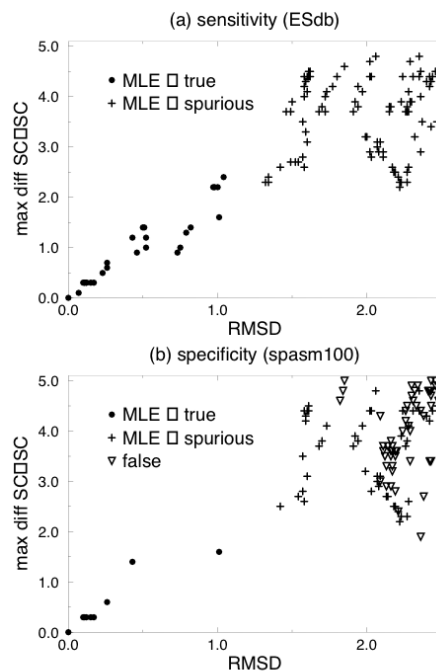
# Distinguishing/clustering subgroups & families

- Subgroup/family searches also work very well with true positives easily separated from other matches, including those of other subgroups/families within the superfamily

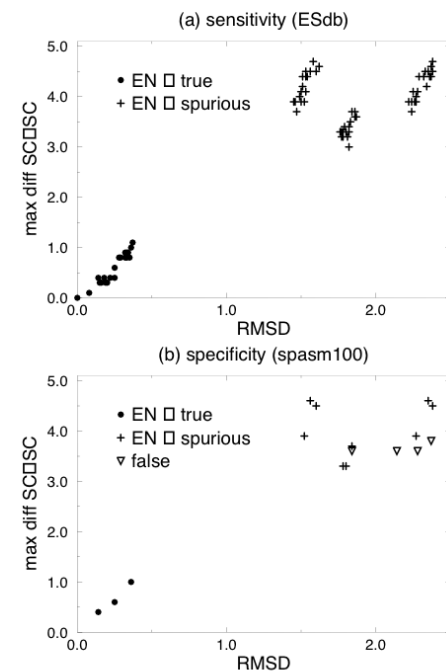
**2mnr MR subgroup template**



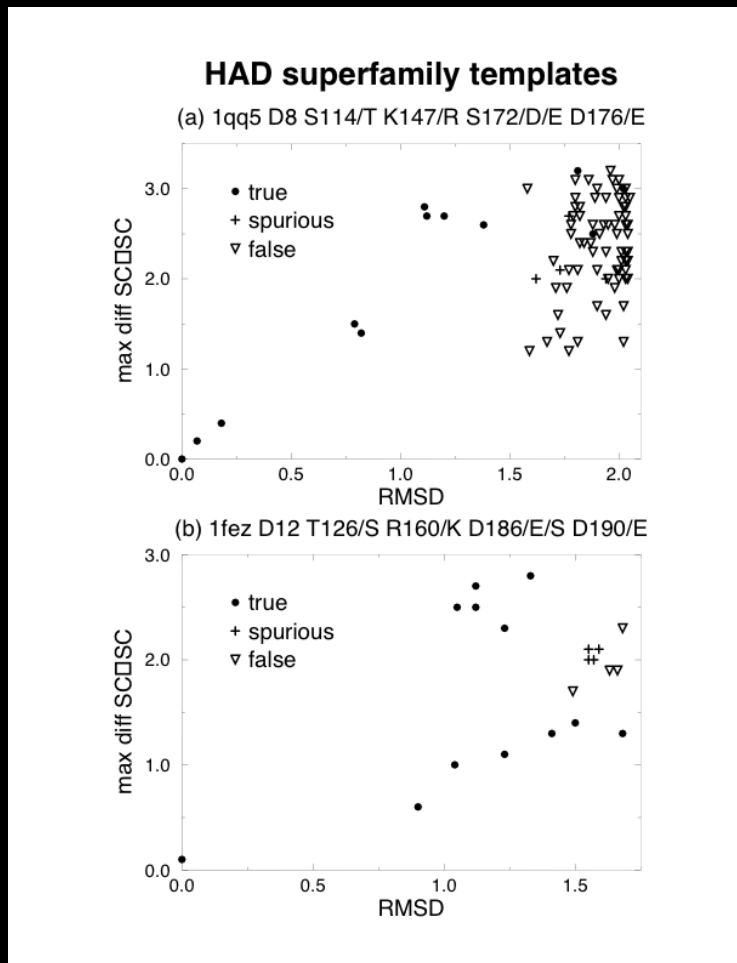
**1muc MLE subgroup template**



**1ebh EN subgroup template**



- A harder problem: Haloacid dehalogenase superfamily (hydrolytic nucleophilic substitution)



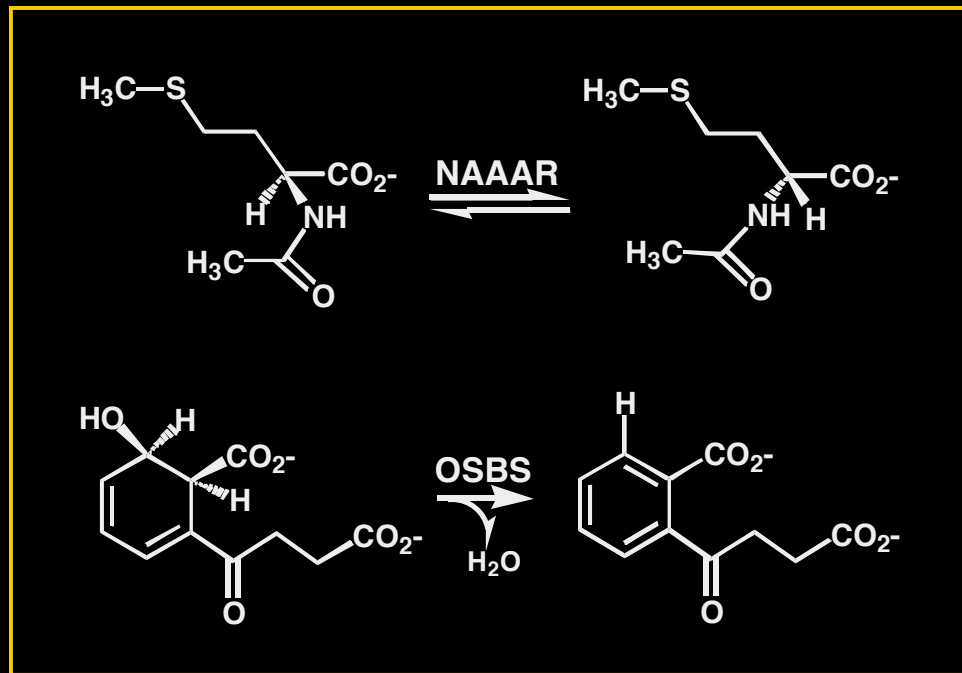
**1qq5: HAD** - to find 12 true positives, must tolerate 4 spurious and 76 false positive matches

**1fez: phosphonatase** - to find 12 true positives, must tolerate 4 spurious and 4 false positive matches



# Protein Engineering

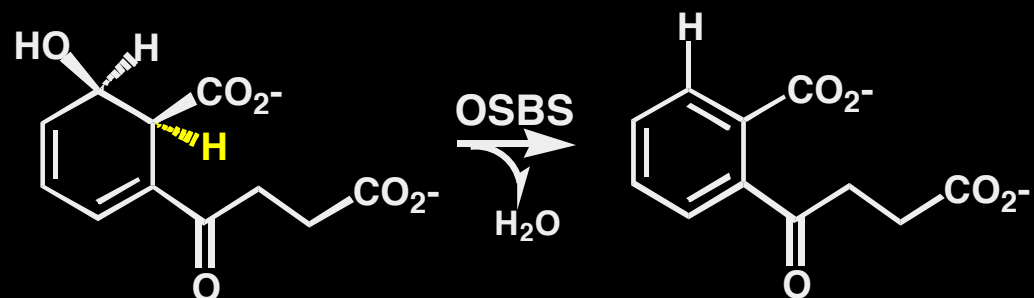
- What template to start with?
- How much difference can we achieve?



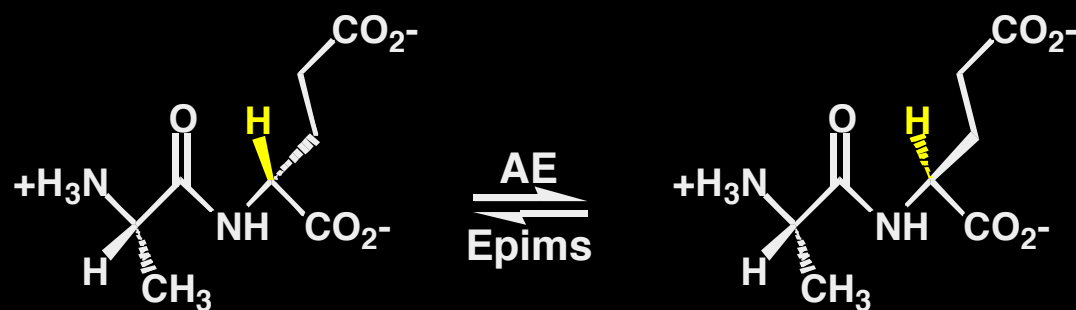
→ Can we take advantage of the conserved chemistry across superfamilies to engineer new reactions?



DNA shuffling



Rational design



Enzyme	$k_{cat}/K_M$ ( $M^{-1}sec^{-1}$ )
<b>OSBS Activity</b>	
OSBS (from <i>E. coli</i> )	$3.1 \times 10^6$
MLE II (wild-type)	$\leq 1.5 \times 10^{-3}$
MLE II E323G	$1.9 \times 10^3$
AEE (wild-type)	$\leq 5.2 \times 10^{-3}$
AEE D297G	12.5
<b>MLE Activity</b>	
MLE II (wild-type)	$2.0 \times 10^4$
MLE II E323G	$1.3 \times 10^3$
AEE (wild-type)	$\leq 5.0 \times 10^{-3}$
AEE D297G	$\leq 6.7 \times 10^{-3}$
<b>AEE Activity</b>	
AEE <i>E. coli</i> (wild-type)	$7.7 \times 10^4$
AEE D297G	9.8
MLE II (wild-type)	$\leq 0.031$
MLE II E323G	$\leq 0.031$

- These results show that the fundamental chemistry associated with the superfamily scaffold can be exploited to generate new overall reactions, including major changes in substrate binding, through single point mutations



# Back to the problem of defining enzyme function

“The analysis of function is tied to the language used to describe it.”

S. Benner (2001) Trends in Genetics 17:414-418

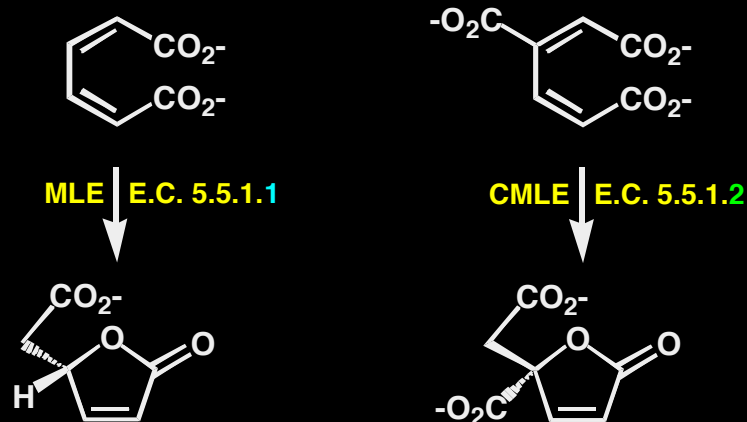
## Why current computationally accessible definitions of function won't work

- For functional inference and related problems, we need structurally contextual definitions of enzyme function that can be manipulated computationally as effectively as we have learned to deal with sequence and structure
  - the available definitions of function as provided by the E.C. system are inadequate because they define function only in terms of overall reactions
  - this won't work for functional annotation using sequence/structural similarities as the starting point because it fails to provide explicit mappings between structure and function

- E.C. may classify structurally similar proteins as functionally dissimilar

<b><i>Superfamily</i></b>	<b><i>Fundamental partial reaction/chemical capability</i></b>
<b><i>Enolase</i></b>	<b><i>Metal dependent abstraction of <math>\alpha</math>-protons of carboxylic acids to form stabilized enolate intermediates</i></b>
EC number	Overall reaction
4.2.1.6	galactonate dehydratase
4.2.1.11	enolase
4.2.1.40	glucarate dehydratase
4.2.1.-	o-succinylbenzoate-CoA synthase
4.3.1.2	methylaspartate ammonia-lyase
5.1.2.3	mandelate racemase
5.5.1.1	muconate lactonizing enzyme
<b><i>Crotonase</i></b>	<b><i>Stabilization of oxyanion intermediates derived from thioesters</i></b>
EC number	Overall reaction
3.1.2.4	3-hydroxyisobutyryl-CoA hydrolase
3.4.21.92	ATP-dependent Clp protease
3.8.1.6	4-chlorobenzoyl-CoA dehalogenase
4.1.1.41	methymalonyl-CoA decarboxylase
4.1.3.36	naphthoate synthase
4.2.1.17	enoyl-CoA hydratase (crotonase)
5.3.3.-	D <sup>3,5</sup> ,D <sup>2,4</sup> -dienoyl-CoA isomerase
<b><i>Haloacid dehalogenase</i></b>	<b><i>Hydrolysis, phosphoryl group transfer via hydrolytic nucleophilic substitution</i></b>
3.1.3.3	phosphoserine phosphatase
3.1.3.15	histidinol phosphatase
3.11.1.1	phosphonatase
3.1.3.18	phosphoglycolate phosphatase
3.8.1.2	haloacid dehalogenase
5.4.2.6	$\beta$ -phosphoglucomutase

- Because E.C. classifications are not associated at any level with structure, annotation transfer between similar sequences on the basis of a known EC number for a database homolog is suspect, especially in mechanistically diverse superfamilies
  - and it is hard to tell what level of similarity is required for any specific superfamily or family
- The converse, inference of structural similarity based on similarities in overall function, e.g., EC # is also problematic
  - includes “analogous” enzymes with the same EC #



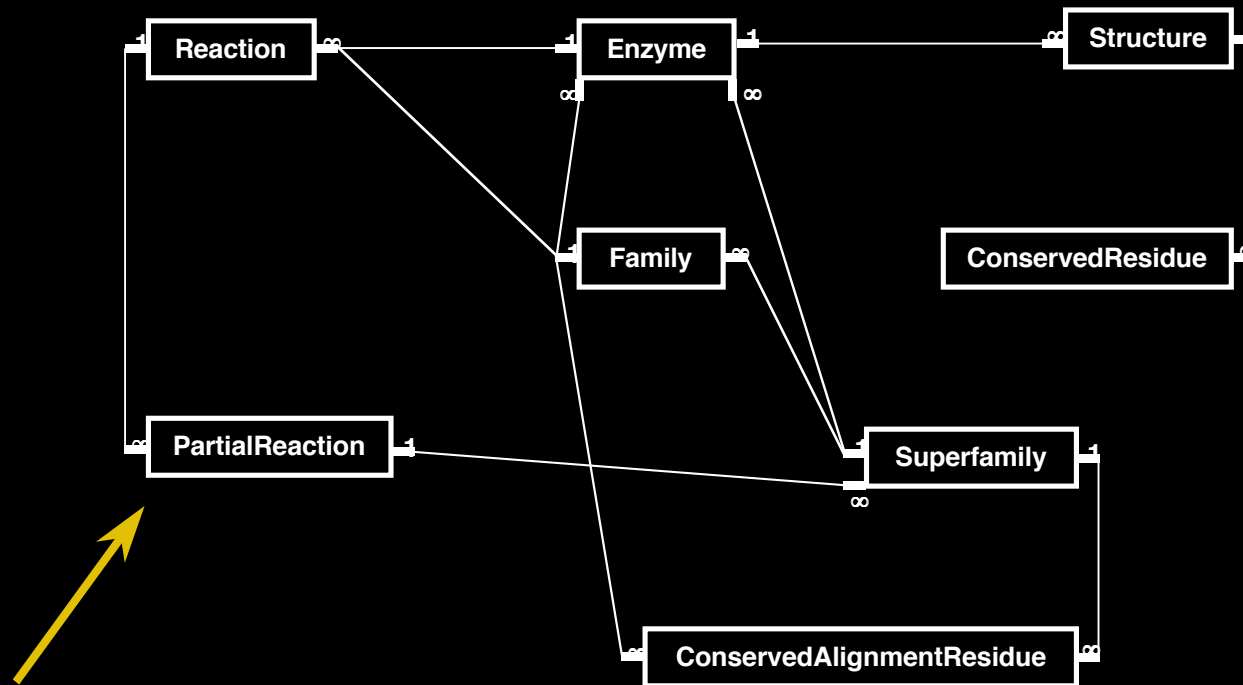
- To address these issues in functional inference for mechanistically diverse enzyme superfamilies, we have created a database for investigation of sequence, structure, and function relationships to aid in functional predictions and the design of protein engineering experiments

(The SFLD is being developed in association with the UCSF Resource for Biocomputing, Visualization and Informatics (RBVI))



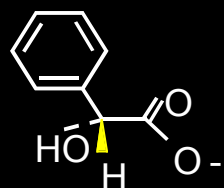
## A new approach to computing with (enzyme) functional information

- Define function not only in terms of the overall reaction but also in terms of the individual partial reactions (or chemical capabilities) that make up each overall reaction

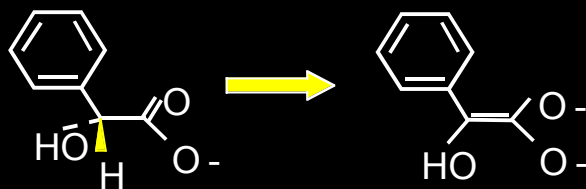


## The Chemical Lexicon

- Description/computation with enzyme chemistry using SMILES/SMARTS



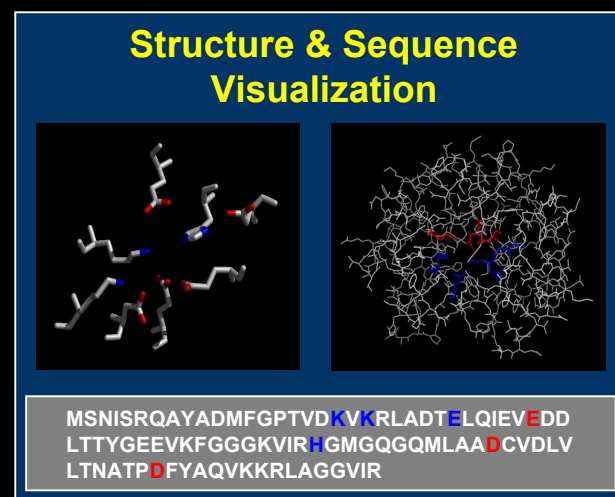
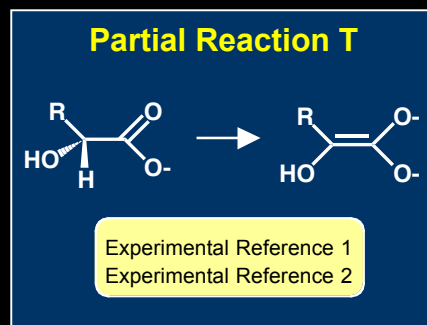
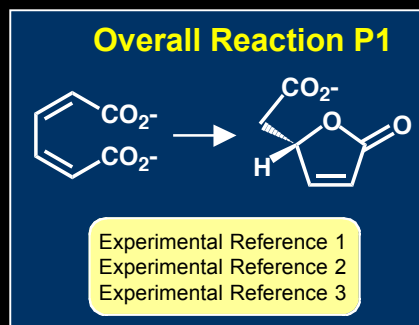
c1ccccc1[C@H](O)C(=O)[O-]



c1ccccc1[C@H](O)C(=O)[O-]>>c1ccccc1C(O)=C([O-])[O-]


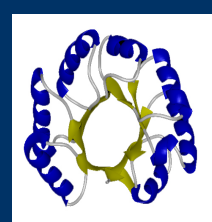

# Applications

- What is the molecular function of my sequence?
  - hypothetical sequences from newly sequenced genomes
  - re-classification of sequences misannotated in the databases
- What protein might be able to perform a function of interest (or that could be engineered to perform that function)?
  - choosing a good template
    - › search by partial reaction
    - › search by substrate
    - › information about diversity of reactions performed by a target superfamily template



- › What are the partial rxns required to achieve the new rxn I want to engineer?
- › Is there a superfamily that “knows” how to do any of these partial rxns?
- › Is there evidence that this superfamily has been used by nature to evolve new functions maintaining my fundamental partial rxn?
- › Will my ligand fit without major rearrangement of the scaffold?

**Search Results**

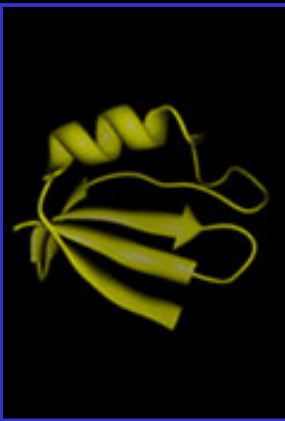
Superfamily A		Superfamily B
Protein 1	Protein 2	Protein 1
		
Partial Reaction F	Partial Reaction T	Partial Reaction J
<b>Partial Reaction T</b>	Partial Reaction B	Partial Reaction J
Partial Reaction U	Partial Reaction U	<b>Partial Reaction T</b>
Partial Reaction Q	Partial Reaction Q	Partial Reaction S
	Partial Reaction R	Partial Reaction T

# SFLD data so far

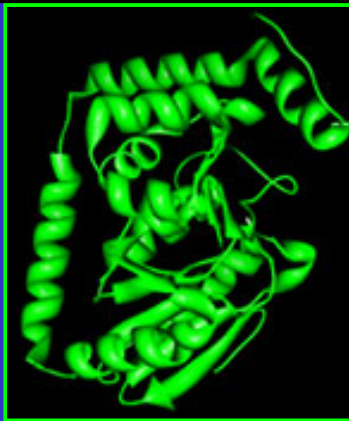
5 Folds, 12 Superfamilies completed or in the pipeline



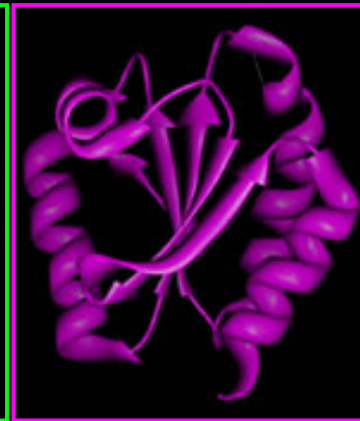
- Enolase SF:  
**457 Proteins**
- Amidohydrolase  
SF: **1736 proteins**
- N-  
acetylneuraminate  
lyase SF: in  
progress



- Vicinal  
Oxygen  
Chelate SF:  
in progress



- Crotonase SF:  
**972 Proteins**



- Haloacid  
Dehalogenase  
SF: **1289  
Proteins**



- Thioredoxin  
SF: in  
progress
  - PDI
  - TRX
  - GRX
  - GST
  - PRX
  - CMP

# Predicting Specificity

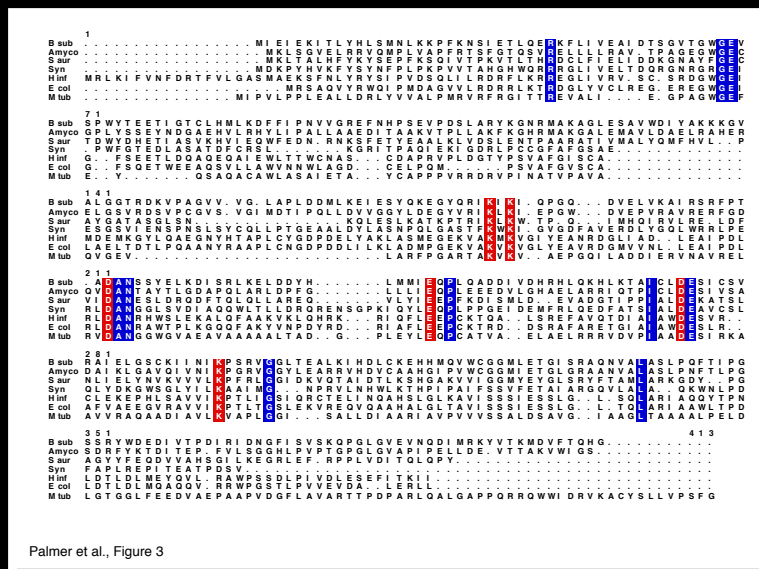
Can we develop predictive rules for specificity in our superfamilies analogous to the rules we have to describe commonality?

- Predicting specificity requires the ability to distinguish families within a superfamily and is especially difficult in mechanistically diverse enzyme superfamilies

## Superfamilies are messy

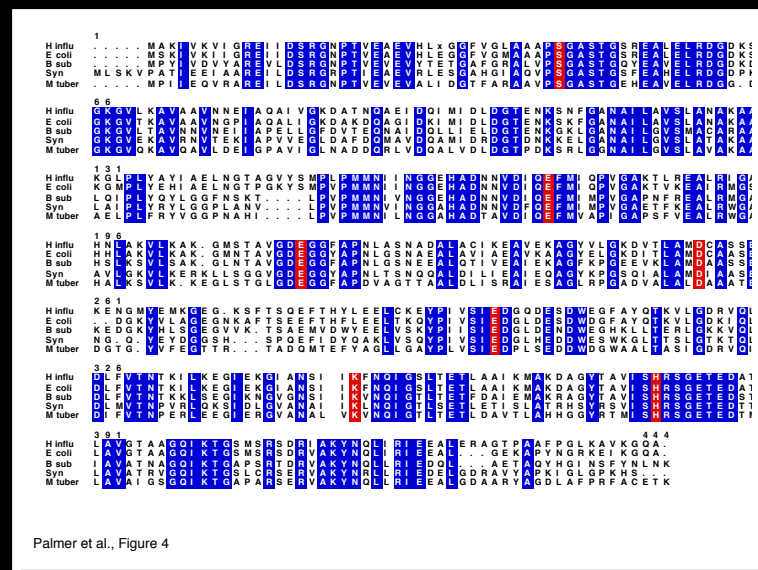
- Relationships are often too distant to easily determine membership in a superfamily
- Connectivity between subgroups and families can be uneven and difficult to evaluate
- Different families/subgroups evolve at different rates
- A given function may have evolved more than once and by different paths
- Distance metrics based on sequence or structure do not track cleanly with divergence of function

- Uneven rates of evolution between families within a superfamily



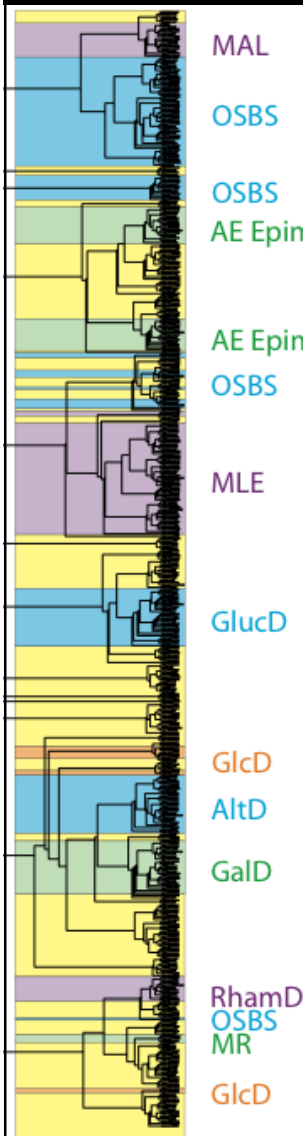
Enolase: 55-86% pairwise identical among a subset of those genomes

OSBS: 14-57% pairwise identical among a set of bacterial genomes





# Convergent evolution within a superfamily: OSBS



Based on operon context and experimental verification, 16 different groups can be identified that appear to perform the OSBS reaction

- these may have arisen via different intermediate ancestors as the superfamily diverged from ancestral genes
- sequences in domains associated with variations in specificity are highly dissimilar

## Coli group

```

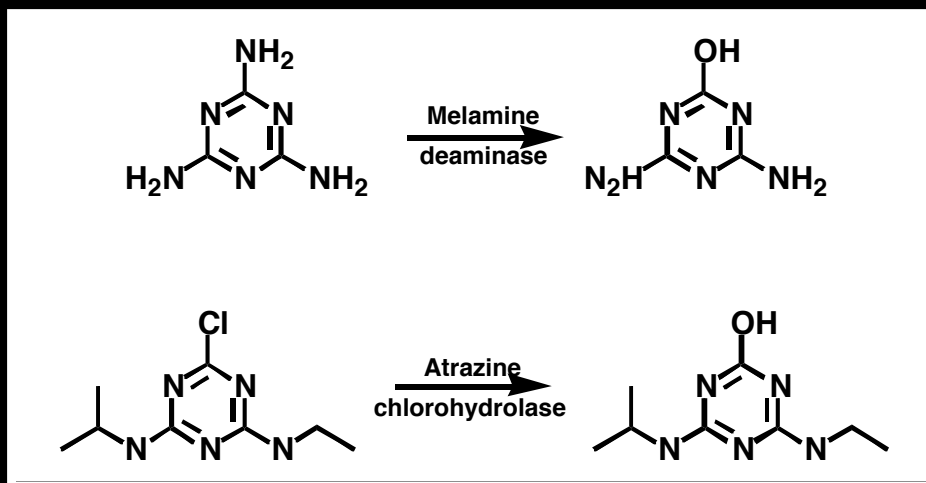
~~~~~Y ---P----- -LR---L--R -G----- ----GE--P LP-FS-E---
~mksakly rytLPmdsgv vLRdekLter vGyielnmn gqkgyGEvsP LPgFSsEtie
~mrhatly ryqlPmdsgv iLRnekLtqR eGfivelten grtarGEiaP LPgFSrEtie
~mrsaqvY rwqiPmdagv vLRdrrLktR dGlyvclrdg eregwGEisP LPgFSqEtwe
~mrsaqvY rwqiPmdagv vLRdrrLktR dGlyvclrdg eregwGEisP LPgFSqEtwe
~mrsaqvY rwqiPmdagv vLRdrrLktR dGlyvclreg eregwGEisP LPgFSqEtwe
~mrsaqvY rwqiPmdagv vLRdrrLktR dGlyvclreg eregwGEisP LPgFSqEnwe
~mrtatly rysvPmeagv iLRhqrLksR dGllvklqqg elsgwGEiaP LPeFSqEtld
~mraatly rysvPmeagv iLRhqrLksR dGllvklqqg eqtgwGEiaP LPeFSqEtld
~mrqailY rysvPmdagv vLRnqrLktR dGllirlhdg eregwGEvaP LPqFSvEtie
~MTnrtfhly qyaiPvdsqL iLRnrfLkkR eGlfvqikcg ehewGEiaP LPeFSqEtie
~MTnrsynly ryaiPvdsqL iLRnrfLkkR eGllvkvcg ehkgwGEiaP LPeFSqEtld
mtmirkfklY qysvPvdsqL iLRnrfLkkR eGllvqvccg daqgwGEiaP LPeFSqEtie
~MaeksfnlY rysvPvdsqL iLRdrfLkR eGlivrvscs .rdgwGEiaP LPgFSqEtld
~mrsakly ryviPvetgt iLRnrrLkqR dGlfqlqdn qrvwGEiaP LPeFSqEtld
~mrkaely ryaiPcqtgv vLRkqpLiqR eGllkleen gkiglGEiaP LPgFSqEtld
  
```

## Amicolaptosis group

```

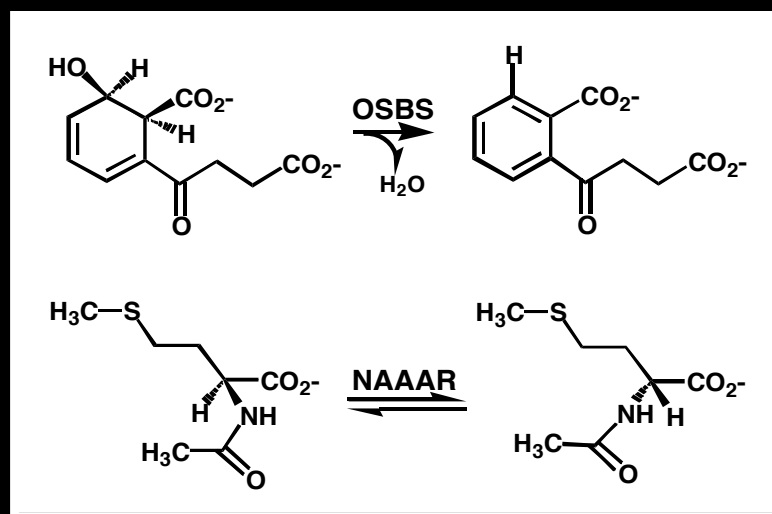
~mklsgvelr rvrmpLvaPF rtSfgtqser elllvravtp aG.eGwGecv ameapLYsse
~mklsgvelr rvrmpLvaPF rtSfgtqser elmlvravtp aG.eGwGecv tmaapvYsse
~mklsgvelr rvqmpLvaPF rtSfgtqsvr ellllravtp aG.eGwGecv tmagpLYsse
~meikkatlh itempLviPF aaSYgtyekr esivielede dGyiGfGEv afePwYteE
~veikkatfh itempLviPF aaSYgtyekr esivielede dGciGfGEv afePwYteE
mieiekitly hlsmnLkkPF knSletlqer kfliveaidt sGvtGwGEvs affspwYteE
~mniqsiety qvrlpLktPF vtSYgrleek afdlfitde qGnqGfGElv afeqPdYvqE
~-----L--PF --S----- -G--G-GE-- ----P-Y--E
  
```

- Highly similar sequences can perform different reactions
  - TriA/AtzA: 98% identical



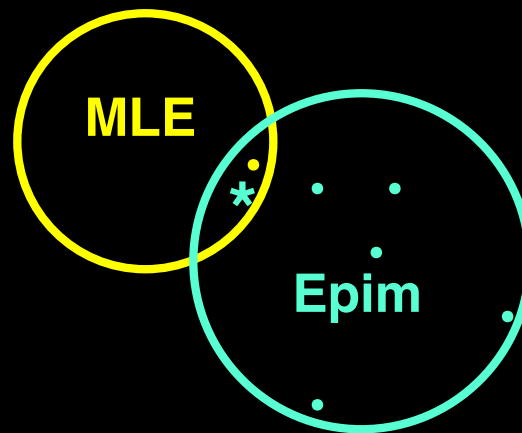
Seffernick et al, J. Bacteriol 183:2405-2410 (2001)

- NAAAR/OSBS: One enzyme performs both rxns



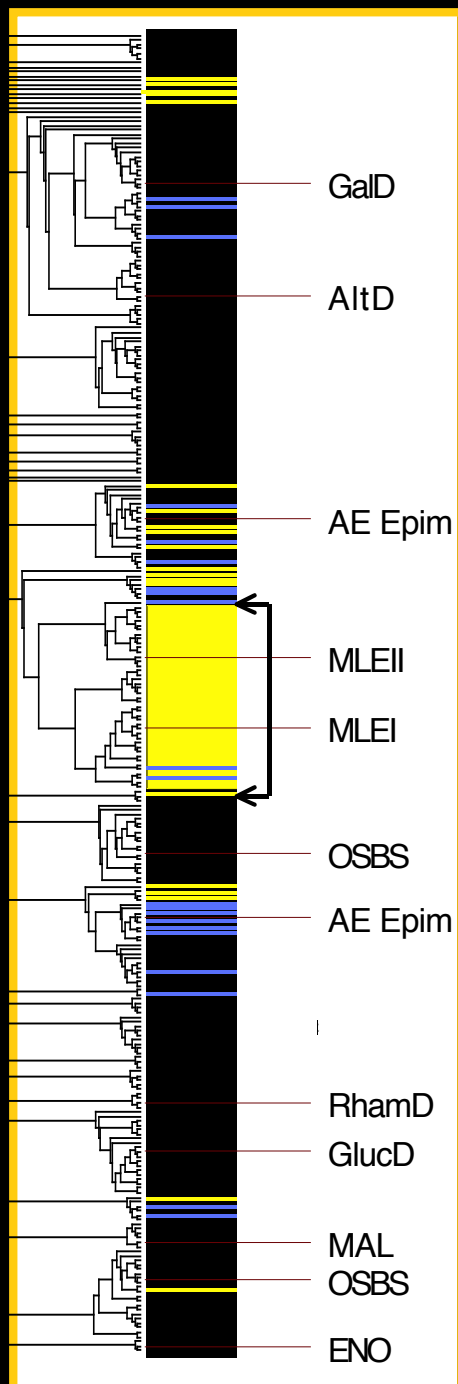
Palmer, D et al., Biochem.38:4252-4258 (1999)

- Sequence-based classification of families performing different functions may have substantial overlap



	Family assignment by BLAST query					
Family	di pep	enolase	galD	mle I	mr	osbs
di pep	7.28E-21	9.00E-02	3.11E-02	1.40E-10	2.00E-02	2.01E-02
enolase	8.18E-02	2.13E-49	9.90E-02	9.29E-02	8.12E-02	9.79E-02
galD	3.22E-15	> 1.00E-01	3.57E-59	6.68E-11	1.62E-16	> 1.00E-01
mle I	5.00E-37	8.17E-02	5.14E-10	4.48E-60	2.33E-15	7.79E-02
mr	2.00E-24	> 1.00E-01	4.15E-17	2.55E-18	5.00E-135	> 1.00E-01
osbs	2.34E-03	9.55E-02	7.27E-02	3.17E-02	5.64E-02	2.52E-02

## One consequence: High levels of misannotation in mechanistically diverse enzyme superfamilies



Sequences Annotated as MLEI from Different Organisms

gi number	Evidence code	HMM	Cat operon
1633162	IDA	1e-273	3
5915882	IDA	4e-238	3
151123	ISO	1e-267	3
23491535	ISO	1e-263	3
4579699	ISO	3e-274	3
7437422	ISO	1e-272	3
5915881	ISO	1e-262	3
9948563	IEA	5e-263	3
13476122	IEA	5e-17	0
23100420	IEA	4e-34	0
15615568	IEA	1e-20	0
15642781	IEA	1e-16	0
17231024	IEA	1e-13	0
29346723	IEA	1e-14	0
23100298	IEA	1e-17	0

IDA: Inferred from direct assay

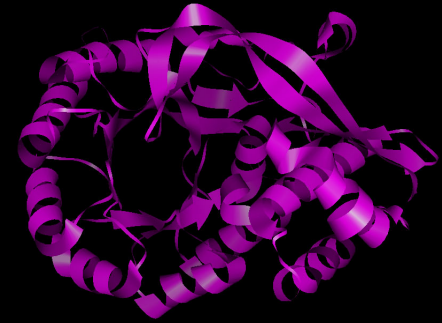
ISO: Inferred from multiple sequence alignments and operon analysis

IEA: Inferred from electronic annotation

- Current estimates in the literature suggest up to 10% of ORFs are misannotated in Genbank with fewer errors in highly curated DB such as SwissProt, COGS
- Our study
  - > 25% MleI and MleII sequences annotated as Mle's are misannotated in Genbank (includes “putatives”)
  - Misannotations are also found for these proteins in SwissProt & Pfam
- How about other superfamilies?
  - Similar levels of misannotation occur in the enoyl CoA hydratase superfamily
  - Work is underway to characterize levels of misannotation in other superfamilies

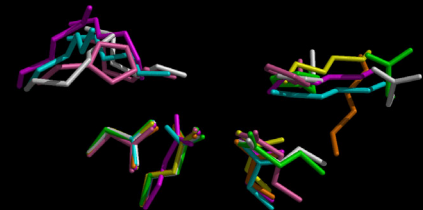
## Functional inference for structural genomics:

What is the function of 1RVK?



Almo, NYSGXRC

- Searching the superfamilies in the SFLD shows it to be a member of the enolase superfamily

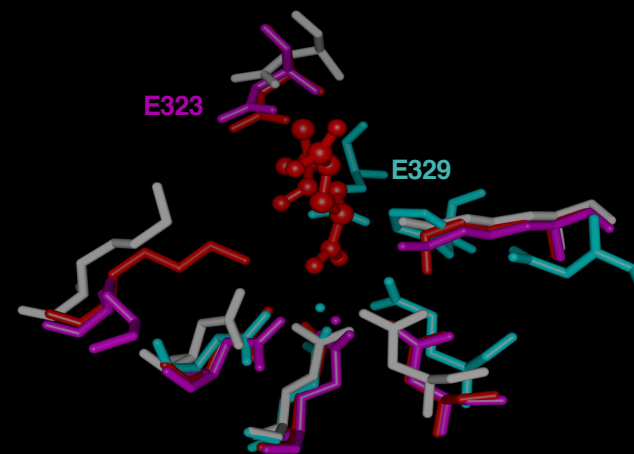


- Currently, the SFLD matches ~5% of ~600 new structures solved by structural genomics projects

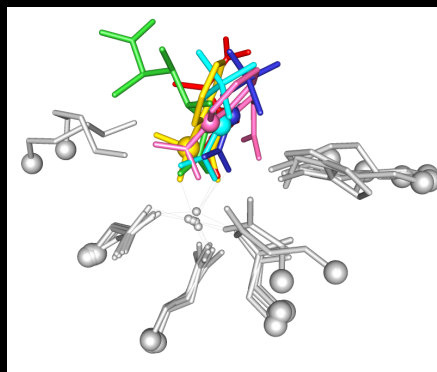
Comparative genomics and operon context suggest 1RVK might be a dipeptide epimerase

However, comparison of an important ligand binding residue (E323) in the D-Ala-L-Glu epimerase from *B. subtilis* with its “homolog” in 1RVK (E329), suggests that the active site may not be big enough for a dipeptide...

The Jacobson group is currently performing docking studies on 1RVK to provide new hypotheses for the identity of the substrate in 1RVK



S-atrolactate ■  
glucarate ■  
methyiaspartate ■  
o-succinylbenzoate ■  
phosphoglycerate ■  
phosphoenolpyruvate ■



Using the superfamily context may aid in filtering docking results for the correct poses and guide the order in which docking solutions should be tested.

# Acknowledgments

## Babbitt Lab

Corey Adams  
Shoshana Brown  
Patty Chang  
Ranyee Chiang  
Margaret Glasner  
Courtney Harper  
Elaine Meng  
Ray Nagatani  
Walter Novak  
Sunil Ojha  
Mark Peterson  
Scott Pegg  
Benjamin Polacco  
Alexandra Schnoes

Jennifer Seffernick (UCSF & Univ. of MN)

**\*John Gerlt (Univ. of Illinois)**

**Shelley Copley (Univ. of Colorado)**

**Larry Wackett (Univ. of MN)**

**Frank Raushel (Texas A&M)**

**UCSF Resource for Biocomputing, Informatics  
& Visualization (RBVI)**

**Tom Ferrin, Director**

**Conrad Huang, Co-Investigator**

**also at UCSF:**

**Andrej Sali**

**Matt Jacobson**

**Brian Shoichet**

\$\$ NIH, NSF, UCSF Sandler Foundation, California Institute for  
Quantitative Biomedical Research (QB3)