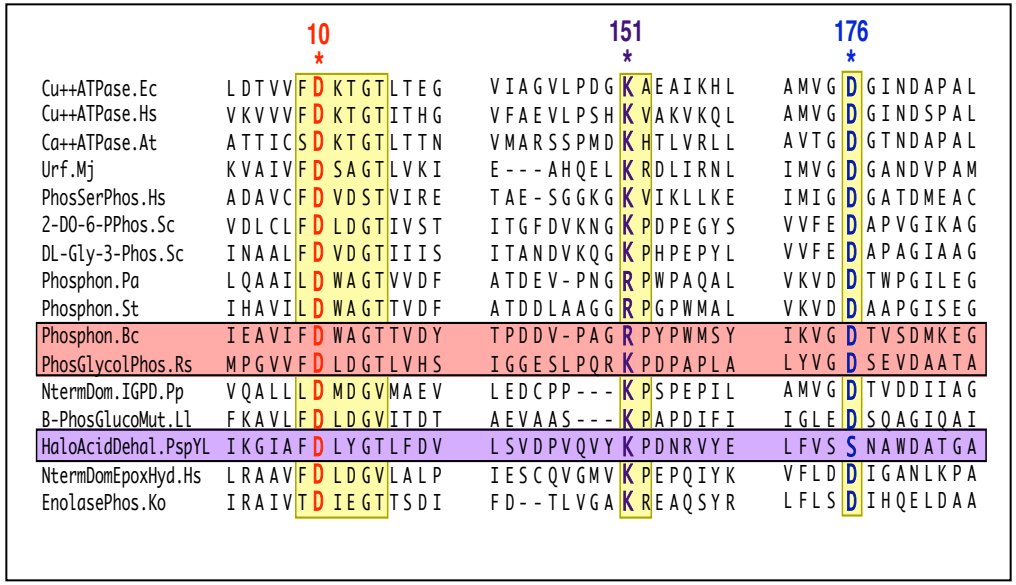
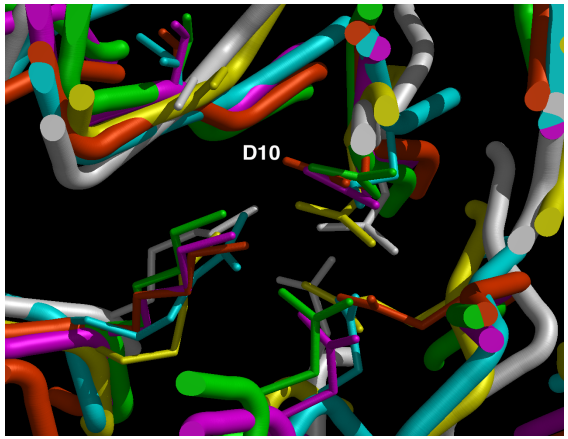
Back to the beginning:
Which sequences to align?

IPAM MSA Workshop
January 2015

Patsy Babbitt
babbitt@cgl.ucsf.edu
University of California, San Francisco

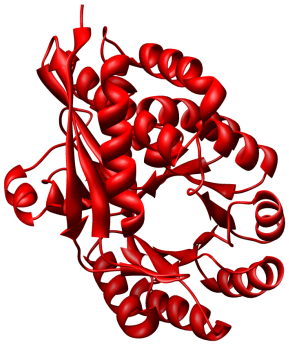
MSAs: Critical for structure-function studies





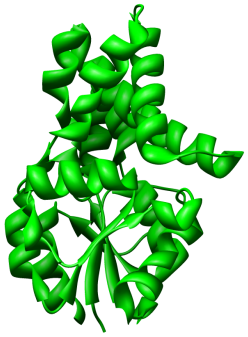
How have enzymes evolved to catalyze the many different chemical reactions required of living organisms?

…by re-using a limited set of “privileged” structural scaffolds

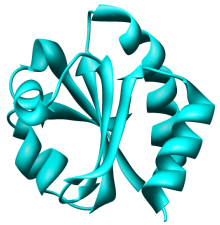


Enolase: 40 rxns

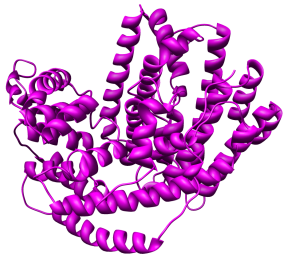
Amidohydrolase: ~100 rxns



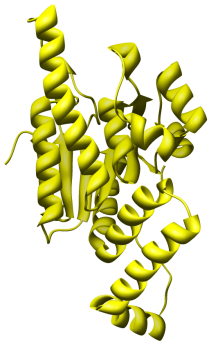
Haloalkanoic acid dehalogenase: 30-50 rxns



Glutathione transferase:44 major subgroups: ?? rxns



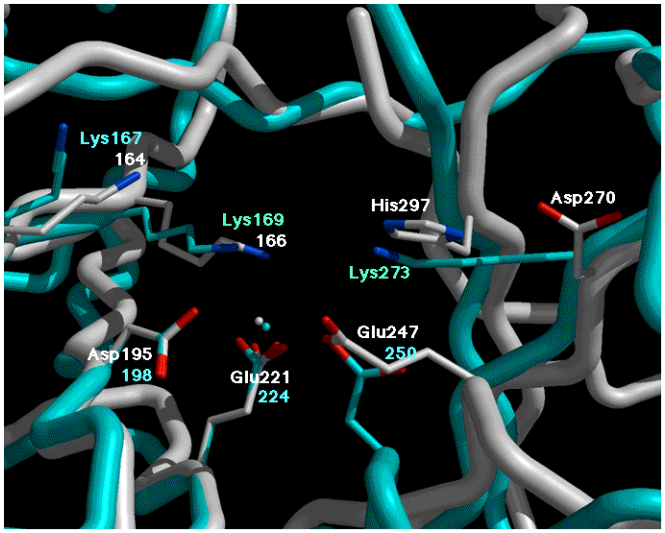
Isoprene synthase I: ?? rxns >50,000 products

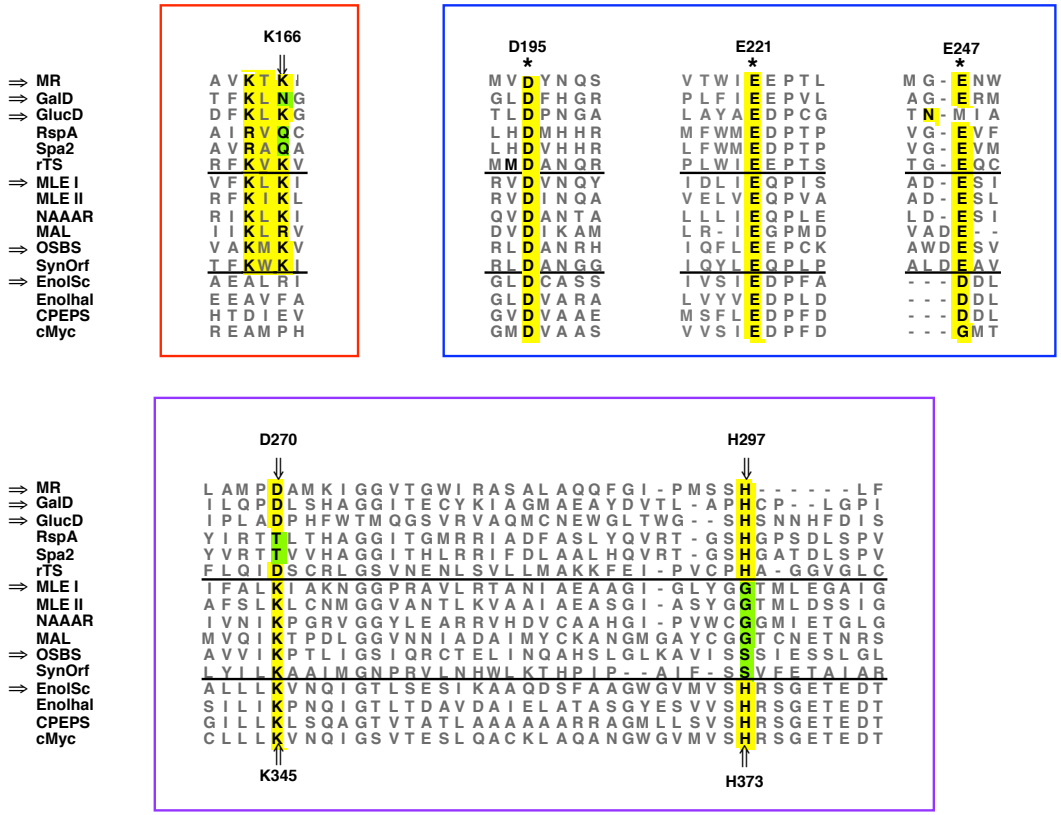


Crotonase: 20 rxns

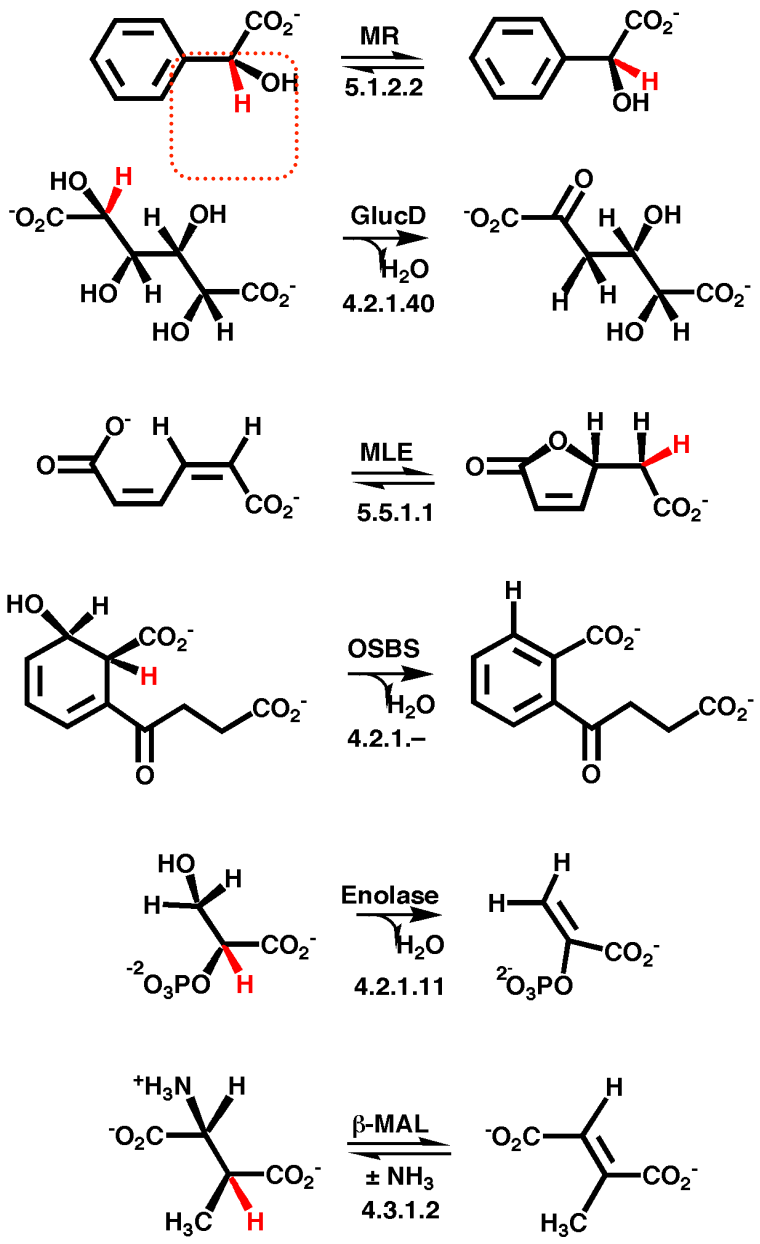
Each functionally diverse enzyme superfamily (SF) links a conserved active site architecture to a fundamental catalytic capability

Example: Enolase SF

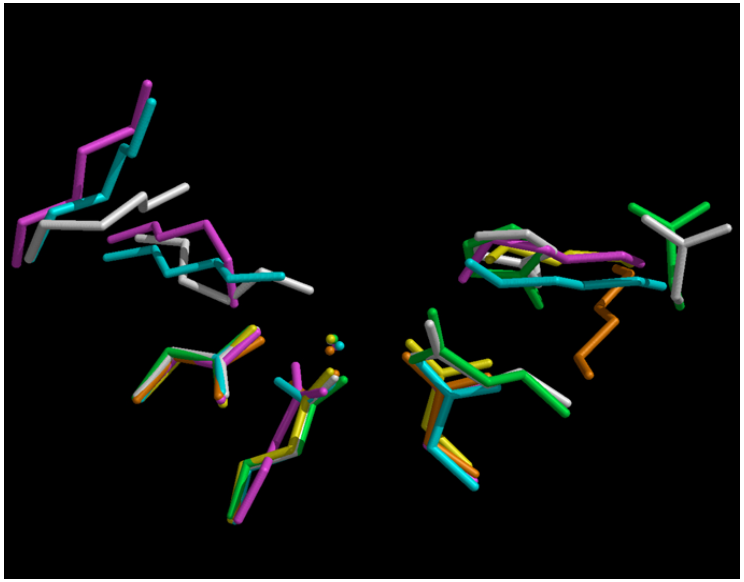




Jensen, R.A. Ann. Rev. Microbiol 30:409-425(1976)
Petsko, G. A. et al., TIBS 18: 372-376 (1993)
Babbitt & Gerlt, JBC 272: 30591 (1997)
Gerlt & Babbitt, Ann Rev Biochem 70: 209 (2001)

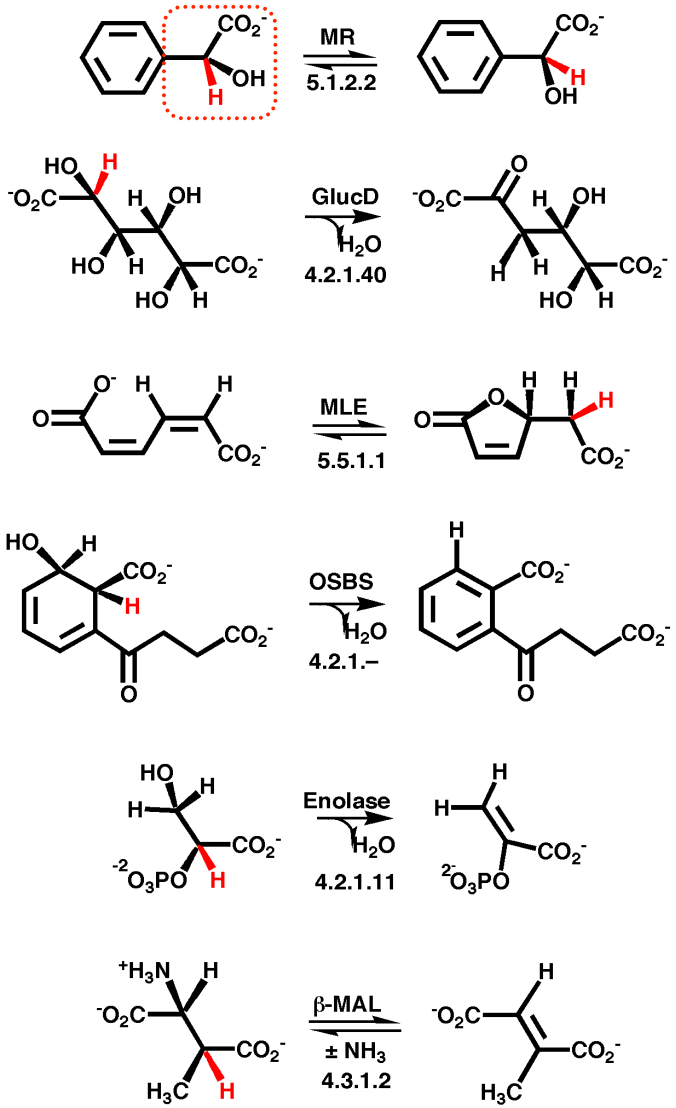


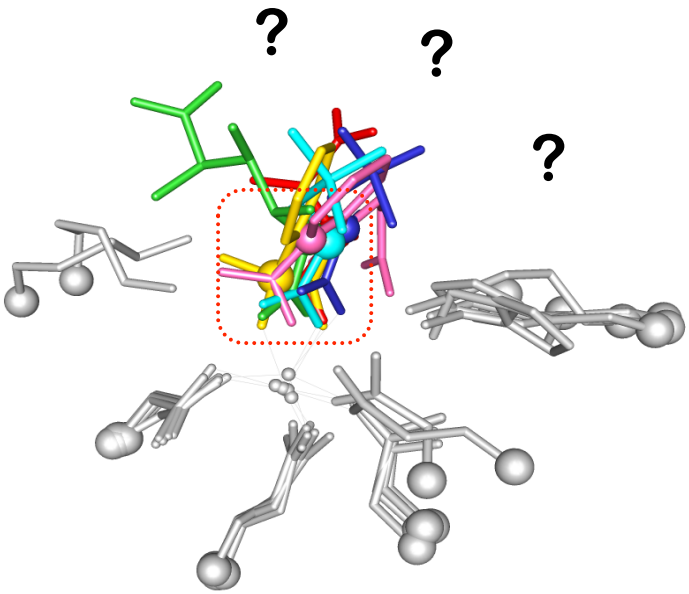
This "chemistry constrained" model restricts the search space for functional inference for all members of the superfamily



>39,000 sequences, 23+ known reactions, many biological functions

Predicting reaction & substrate specificity is much harder …





Goals

- Identify the fundamental chemical capability associated with the conserved features of the structural superfamily and the catalytic machinery shared by all members
- Determine the patterns by which divergence has altered the structural scaffold and active site architecture to enable many different chemical reactions using the superfamily scaffold

Applications

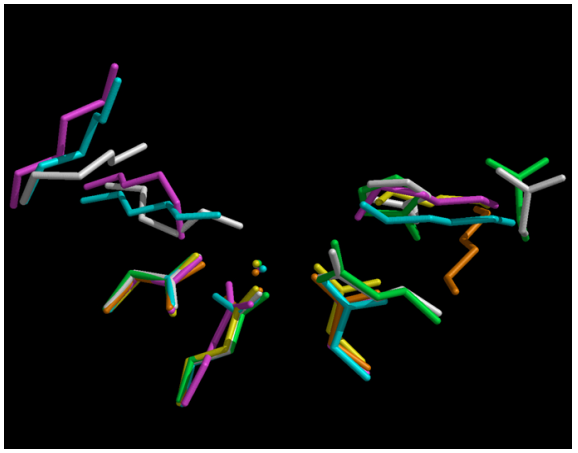
- Inform mechanism in knowns
- Use a SF's fundamental structure-function mapping to limit the search space for functional inference in unknowns
- Curation of functionally diverse enzyme superfamilies on a large scale
- Guide identification of starting scaffolds for enzyme engineering
  - which superfamily scaffolds "know" how to catalyze the thermodynamically difficult step for a reaction of interest?

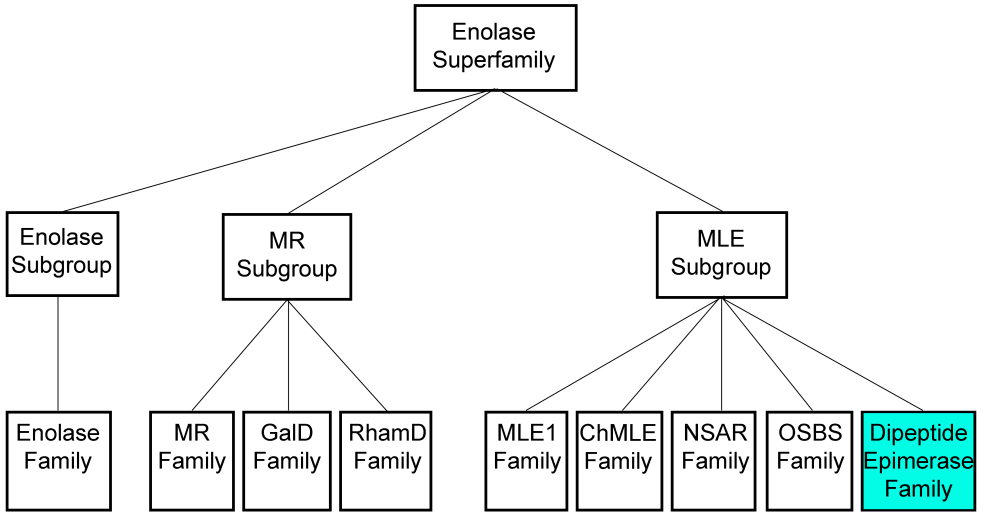
What does this have to do with MSA creation?

- How to take advantage of the volume of [diverse] sequence and structure data now available
- How to incorporate more functionally/biologically informed ways to choose which sequences/structures to align

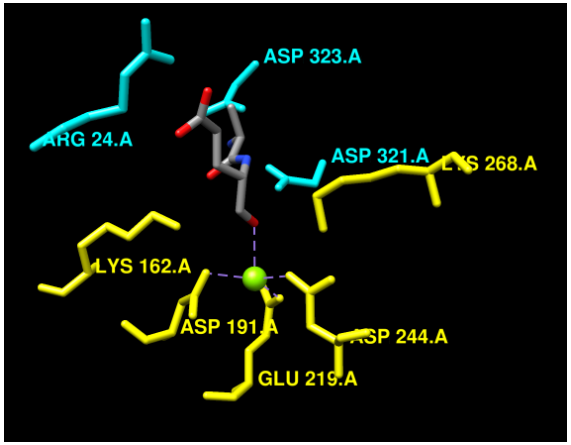
This chemistry-constrained model suggests a general classification scheme

Superfamily: Subset of active site residues conserved in all members of a superfamily catalyze a conserved partial rxn



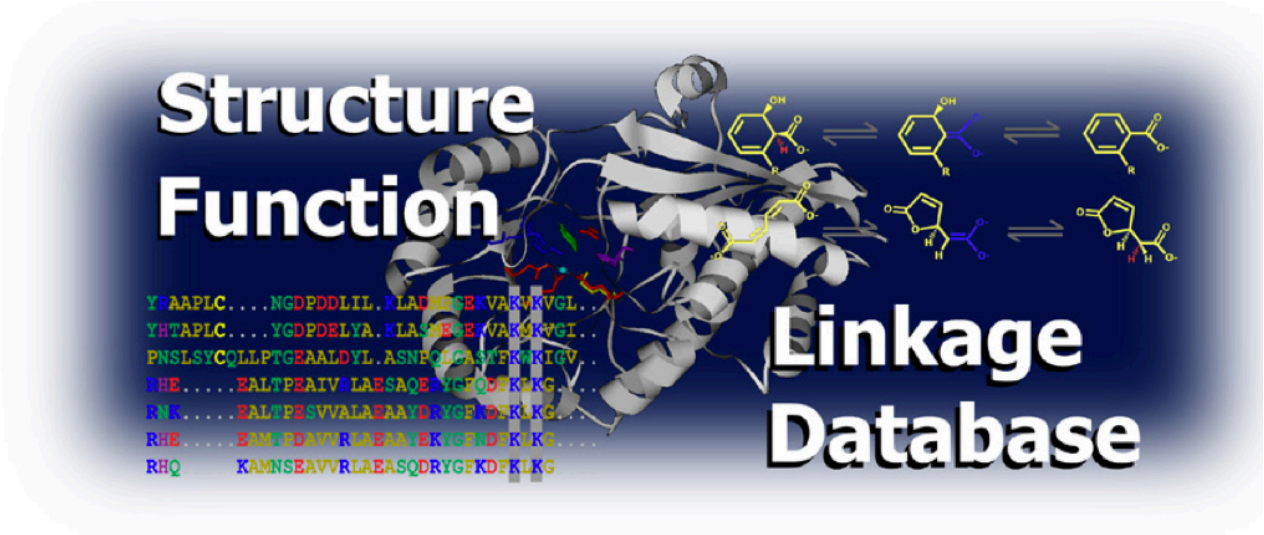


Family: Family-specific subsets of residues mediate reaction & substrate specificity



Annotation transfer only at the level of granularity at which good supporting evidence exists

SFLD: http://sfld.rbvi.ucsf.edu



See description in ***Nucleic Acids Res.*** 2014 Jan 1;**42**:D521-30

What is the Structure-Function Linkage Database (SFLD)?

- A hierarchical classification of enzymes that relates specific sequence-structure features to specific chemical capabilities
- A collection of tools and data for investigating sequence-structure-function relationships and hypothesizing function
- More...

How can I use the SFLD?

(see the tutorials for examples)

- **Classify a sequence** using Hidden Markov Models or BLAST search
- **Browse superfamilies in the SFLD**
- **Browse reactions** (overall)
- **Search for a specific enzyme** (by name, sequence database ID, or PDB ID)
- **View sequence alignments**
- **View structures in Chimera**
- **Download data**: sequence sets, multiple alignments, sequence similarity networks...

What makes the SFLD unique?

- Superfamilies are defined by a conserved chemical capability such as a partial reaction, families by a conserved overall reaction (more...)
- Conserved partial reactions are correlated with associated active site similarities
- Large-scale summaries of relationships between and within groups of enzymes are provided as sequence similarity networks

Projects under Development

- Extended SFLD (XSFLD)
- Identifying Potential Misannotations

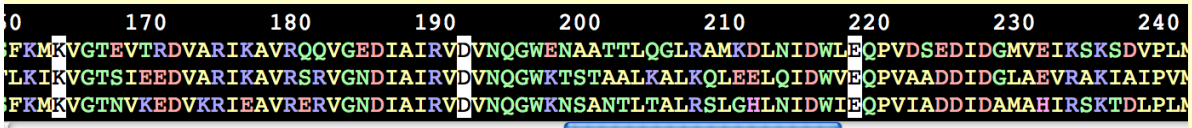
Top Level	Name
└ Superfamily (core)	Enolase
└ Subgroup	muconate cycloisomerase
└ Family	N-succinylamino acid racemase 2

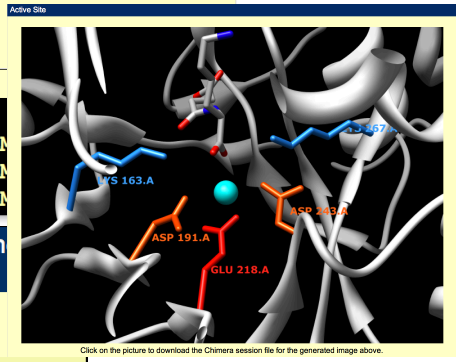
2P8B (N-succinylamino acid racemase 2)

	Total	100%	<100%
Functional domains	196	0	196
UniProtKB	255	0	255
GI	694	0	694
Structures	3		
Reactions	1		

Functional domains of this family were last updated on Jan. 7, 2015
New functional domains were last added to this family on Dec. 31, 2014

Pairwise % Identities



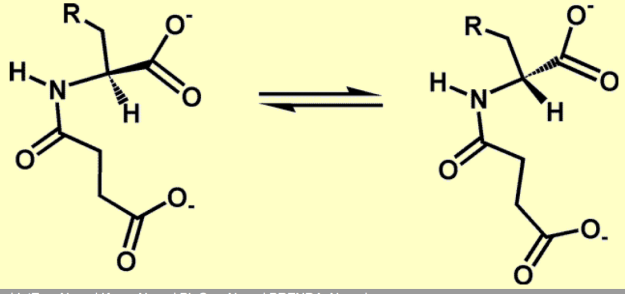


Click on the picture to download the Chimera session file for the generated image above.

Position	Amino acid	Function	Curator notes	Eviden Code
163	Lys (K)	base (abstracts alpha proton), acid (donates proton to leaving group)		ICS
191	Asp (D)	metal binding ligand		ICS
218	Glu (E)	metal binding ligand		ICS
243	Asp (D)	metal binding ligand		ICS
267	Lys (K)	base (abstracts alpha proton), acid (donates proton to leaving group)		ICS

Catalyzed Reaction(s)

racemization of n-succinylamino acid



C: None | IntEnz: None | Kegg: None | BioCyc: None | BRENDA: None |

Protein Name	Superfamily	Family	Species	Databases	UniProtKB	MicrobesOnline Operon	The SEED	PDB ID	Structures	Updated
N-succinylamino acid racemase 2	Enolase	N-succinylamino acid racemase 2	Bacillus anthracis str. Ames	GI REF UP MO TS MB PDB	A0A084CRN0	7611946	fig|486623.3.peg.5055	2P8B 2P8C 2P88	3	Jan. 07, 2015
N-succinylamino acid racemase 2	Enolase	N-succinylamino acid racemase 2	Bacillus cereus ATCC 14579	GI REF UP MO TS MB PDB	Q81IL5	357748	fig|226900.1.peg.322	2P8B 2P8C 2P88	3	Jan. 07, 2015
N-succinylamino acid racemase 2	Enolase	N-succinylamino acid racemase 2	Bacillus cereus ATCC 10987	GI REF UP MO TS MB PDB	Q73EC5	643701	fig|222523.1.peg.433	2P8B 2P8C 2P88	3	Jan. 07, 2015
N-succinylamino acid racemase 2	Enolase	N-succinylamino acid racemase 2	Bacillus cereus	GI REF UP TS MB PDB	Q4MJ00		fig|269801.1.peg.4979	2P8B 2P8C 2P88	3	Jan. 07, 2015
N-succinylamino acid racemase 2	Enolase	N-succinylamino acid racemase 2	Bacillus thuringiensis	GI REF UP MO MB PDB	Q6HP62	595563		2P8B 2P8C 2P88	3	Jan. 07, 2015

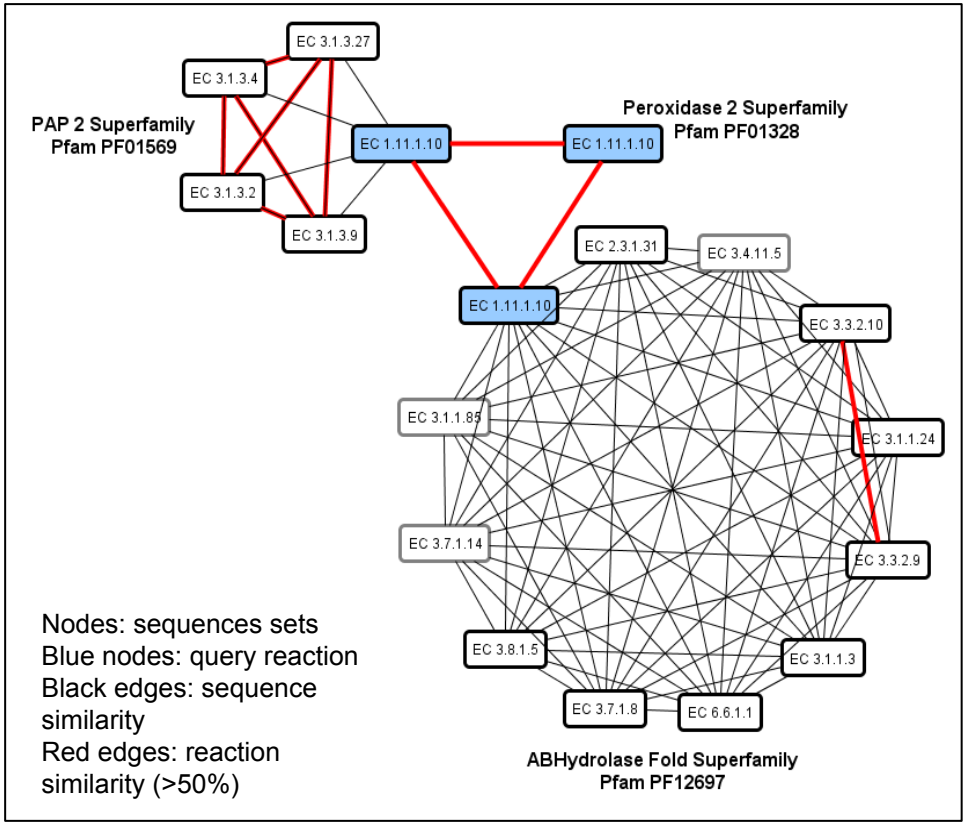
Core SFLD

Suprafamily	Superfamily	Subgroups	Families	Sequences	Structures	Reactions
	Amidohydrolase	11	89	79238	474	41
	Aromatic Prenyltransferase	2	0	339	18	0
	Crotonase	2	27	75290	172	28
	Enolase	7	20	39661	355	22
	Haloacid Dehalogenase	25	22	79778	570	21
	Isoprenoid Synthase Type I	14	69	16579	359	65
	Isoprenoid Synthase Type II	4	8	7645	202	8
	Nucleophilic Attack 6-Bladed Beta-Propeller (N6P)	3	3	31085	85	2
	Radical SAM	49	93	113568	52	66
	RuBisCO	2	2	41212	73	2
Thioredoxin Fold Suprafamily	Glutathione Transferase (cytosolic)	42	0	13097	432	0
	Peroxiredoxin	6	0	12239	179	0
	TOTAL: 12	167	333	515791	2971	255

Next steps: Identification & integration of conserved chemical capabilities in superfamilies & patterns by which their chemistry varies

Superfamily	SCOP ID	Number of Reactions	Number of Unique EC numbers	Conserved EC positions	Conserved Substructure
Alkaline phosphatase-like	c.76.1	67	5	3.1.x.x	OP(=O)O
SGNH hydrolase	c.23.10	19	2	3.1.1.x	CC=O
Metallo-dependent phosphatases	d.159.1	30	2	3.1.3.x	COP(O)(=O)O
Carbohydrate phosphatase	e.7.1	17	4	3.1.3.x	CCOP(=O)(O)O
Cobalamin (vitamin B12)-dependent enzymes	c.1.19	6	2	4.2.1.x	OCCO
Phosphoglycerate mutase-like	c.60.1	58	5	x.x.x.x	COP(O)(O)=O
Six-hairpin glycosidases	a.102.1	6	3	x.x.x.x	OCC1OC(O)CC(O)C1(O)
alpha/beta-Hydrolases	c.69.1	13	3	3.1.1.x	CCOC(C)=O

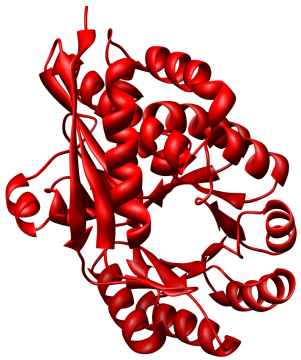
Chiang et al, PLoS CB, 4:e1000142 (2008)



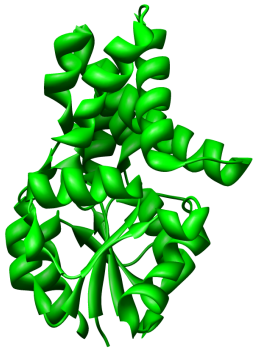
Gemma Holliday PhD

Tackling sequence data on a large scale

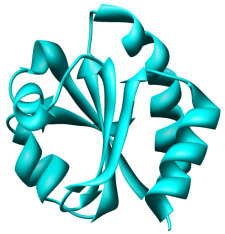
Most protein superfamilies are too large to manage & easily explore using multiple sequence alignments & trees



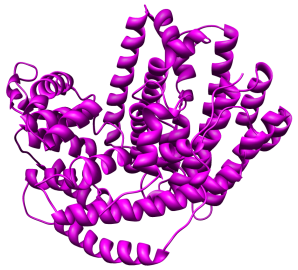
Enolase
~39,000 seqs



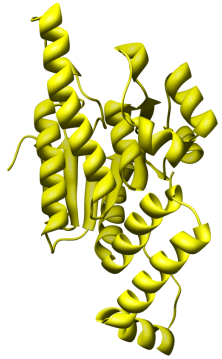
Haloalkanoic acid dehalogenase
>80,000



Glutathione transferase
>50,000



Isoprene synthase I
>>16,000

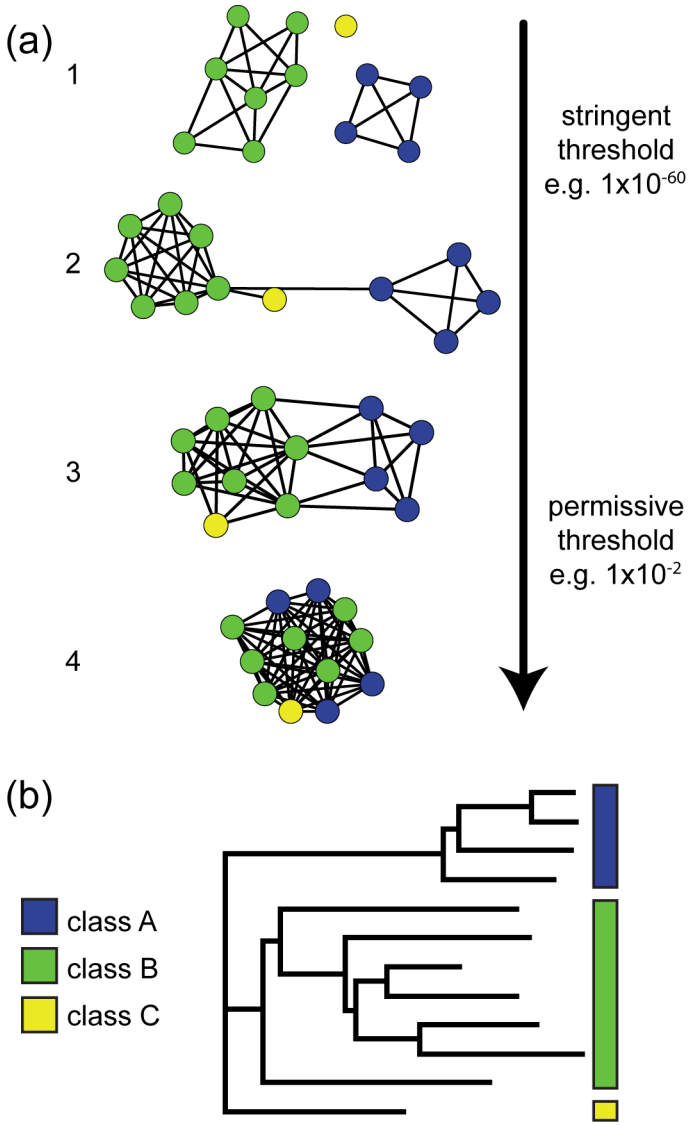


Enoyl CoA hydratase
>72,000

We estimate that ~1/3 of the universe of enzyme superfamilies are functionally diverse

Almonacid & Babbitt, Curr Op Biol Chem 15:435 (2011)

Accessing the larger context: Protein similarity networks



- > Powerful hypothesis generator for structure-function relationships
- > Interactive
- > Fast
- > Easy visualization (Cytoscape)
- > Handles thousands of sequences, structures, etc.
- > Pairwise comparisons don't require a multiple alignment
- > Track well with known structure-function relationships
- > Capture sequence, structure, ligand, relationships
- > Many metrics & algorithms for comparison & clustering of similarity data
- > Networks not a substitute for phylogenetic trees

Node = sequence (or structure)

Edge = connections between sequences w scores as good as the E-value cutoff threhold

Sequence networks in this presentation: all-by-all BLAST (*E*-values as scores); structure: all-by-all FAST/TM-align scores

Atkinson HJ & Babbitt PC. *PLoS ONE* (2009) 4: e4345

Validation

Comparison of distance metrics for generating networks shows that BLAST correlates well with several other metrics

Uronatel / **MLE**	BLAST	SW	MA	PT
BLAST		0.999	0.971	0.953
Smith-Waterman (SW)	0.998		0.970	0.953
Multiple Alignment (MA)	0.800	0.798		0.974
Phylogenetic Tree (PT)	0.731	0.731	0.777	

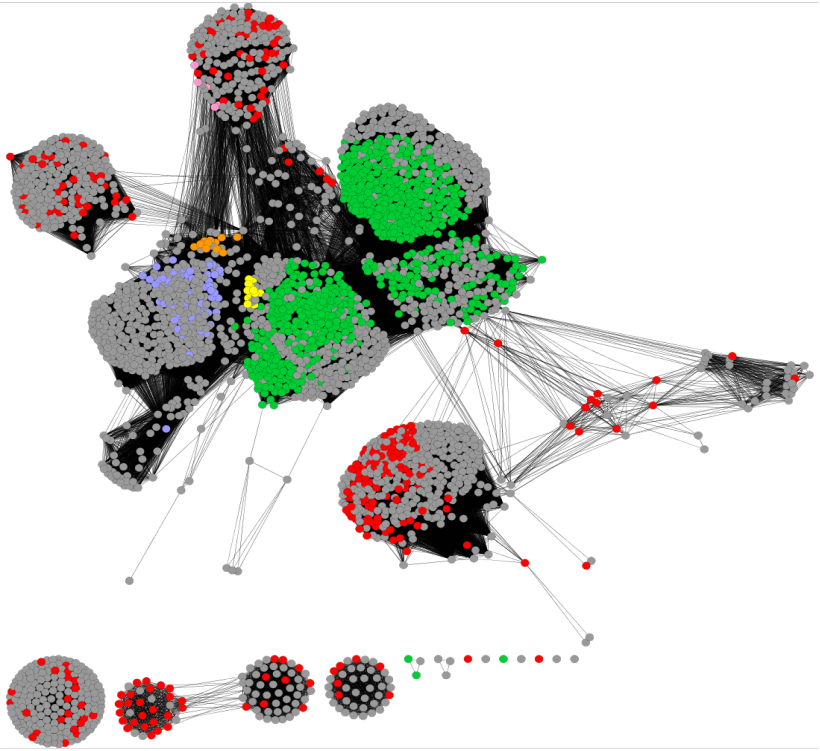
NagA	BLAST	SW	MA	PT
BLAST		0.997	0.841	0.748
Smith-Waterman (SW)			0.846	0.753
Multiple Alignment (MA)				0.719
Phylogenetic Tree (PT)				

R^2 values for linear regressions of distances generated from various metrics for scoring similarity among sequences

Other validation analyses show

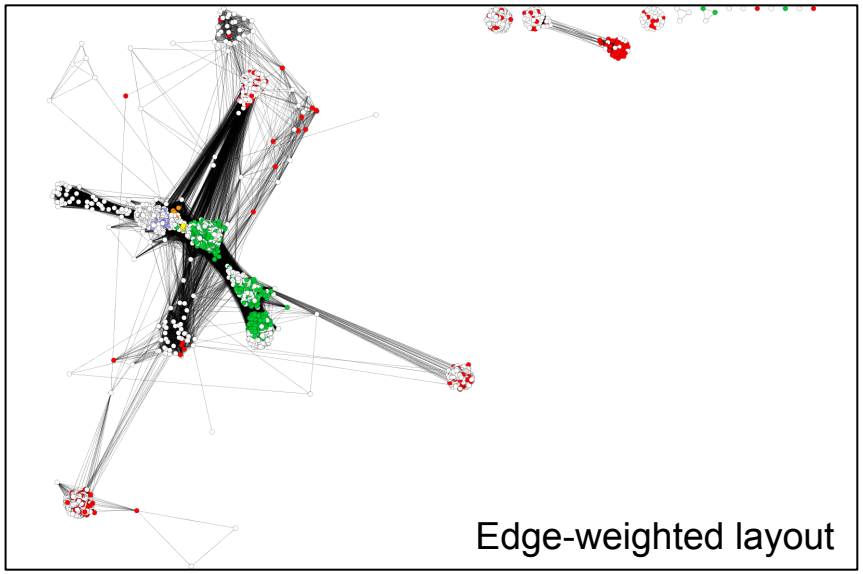
- Network topologies are generally robust to missing data
- Two-dimensional distances in visualized networks correlate well with the underlying distances in high-dimensional space
- See Atkinson et al, PLoS ONE, 4: e4345 (2009) for more statistical validation of PSNs

Layout used in this talk



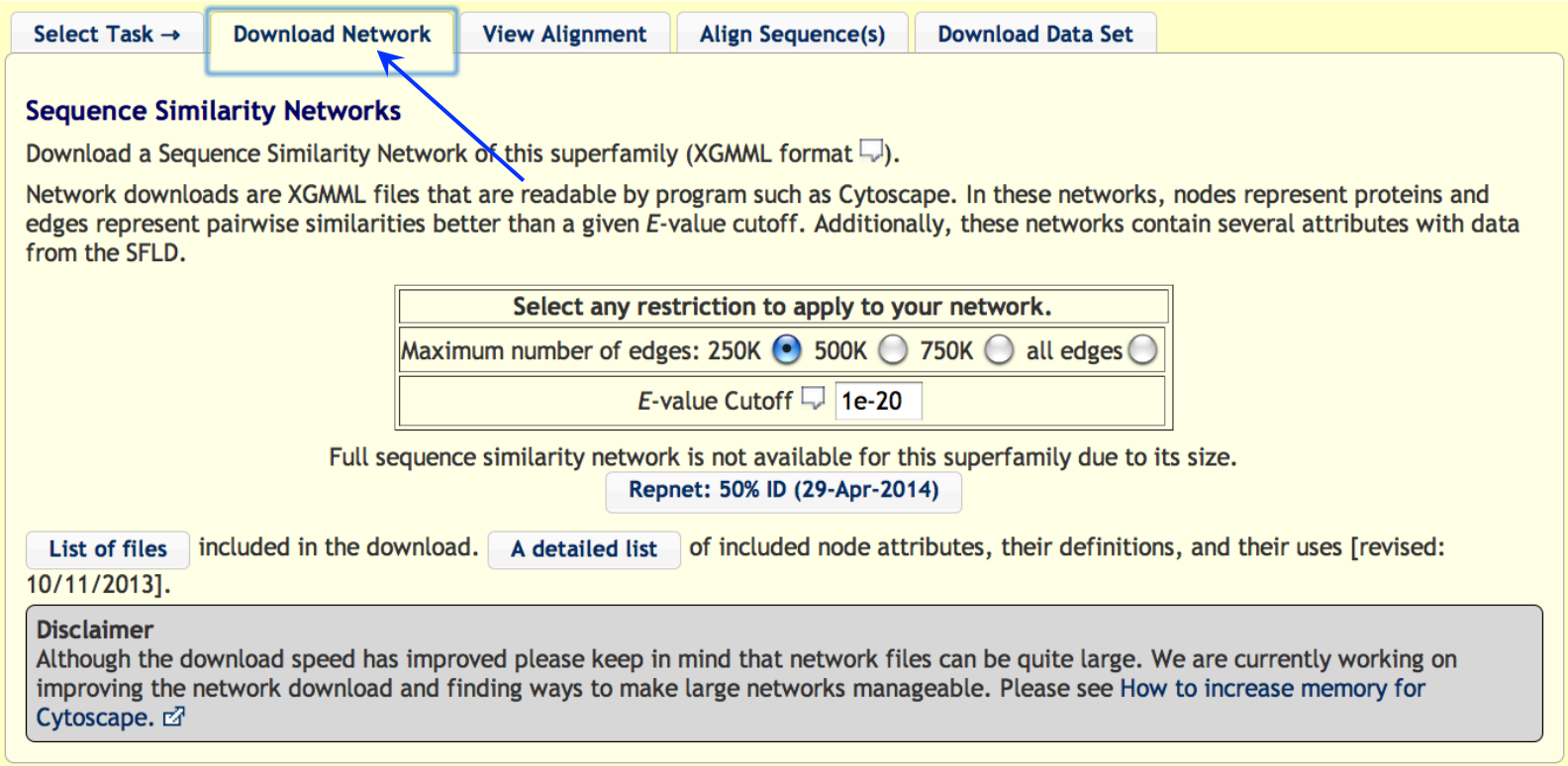
Organic layout: Edge lengths represent degree of connectivity, track with dissimilarity

Colors depict function or other types of functional information



Suprafamily	Superfamily	Subgroups	Families	Sequences	Structures	Reactions
	Amidohydrolase	11	89	79238	474	41
	Aromatic Prenyltransferase	2	0	339	18	0
	Crotonase	2	27	75290	172	28
	Enolase	7	20	39661	355	22
	Haloacid Dehalogenase	25	22	79778	570	21
	Isoprenoid Synthase Type I	14	69	16579	359	65
	Isoprenoid Synthase Type II	4	8	7645	202	8
	Nucleophilic Attack 6-Bladed Beta-Propeller (N6P)	3	3	31085	85	2
	Radical SAM	49	93	113568	52	66
	RuBisCO	2	2	41212	73	2
Thioredoxin Fold Suprafamily	Glutathione Transferase (cytosolic)	42	0	13097	432	0
	Peroxiredoxin	6	0	12239	179	0
	TOTAL: 12	167	333	515791	2971	255

For the Core SFLD, networks can be downloaded at every level of the hierarchy: Superfamily, Subgroup, & Family



Extended SFLD

Suprafamily	Superfamily	Subgroups	Families	Sequences	Structures	Reactions
	Arginase/Deacetylase	0	0	10570	177	0
	Carbohydrate Phosphatase	0	0	12278	160	0
	Carbon-Nitrogen Hydrolase	0	0	14974	49	0
	Chelatase	0	0	5776	49	0
	Cytidine Deaminase-Like	0	0	18803	90	0
	Di-trans-poly-cis-decaprenylcistransferase	0	0	4075	36	0
	dUTPase-Like	0	0	6343	156	0
	Ferric Reductase Domain	2	2	282	1	0
	Fumarylacetoacetase, C-terminal-related	0	0	10104	32	0
	Glutaminase/Asparaginase	0	0	2672	37	0
	HD-Domain/PDEase-Like	0	0	33746	252	0
	Histidine Phosphatase	0	0	21228	254	0
	Isochorismatase-Like Hydrolases	0	0	11412	42	0
	Kringle-Like	0	0	1693	627	0
	L-Aspartase-Like	0	0	17659	100	0
	Metalloproteases, Zincins	0	0	21437	429	0
	Methyltransferase Domain 18	0	0	4312	42	0
	Methyltransferase Domain 9	0	0	433	1	0
	NUDIX Hydrolase Domain-Like	0	0	38313	230	0
	Peptidase M24	0	0	17752	177	0
	Phosphatidylinositol Phosphodiesterase	3	5	11014	97	5
	Phospholipase C/P1 Nuclease	0	0	1560	21	0
	Phosphonate Radical SAM	1	1	902	0	1
	PLP-Binding Barrel	0	0	19071	95	0
	Proline Racemase	0	0	748	14	0
	Pyruvoyl-Dependent Histidine/Arginine Decarboxylase	2	2	335	11	2
	Radical SAM 3-amino-3-carboxypropyl Radical Forming	1	1	1644	2	1
	Ribulose-Phosphate Binding Barrel	0	0	25997	249	0
	SGNH Hydrolase	0	0	19406	63	0
	Six-Hairpin Glycosidases	0	0	28690	227	0
	SPOUT Methyltransferase	0	0	2593	5	0
	Subtilisin-Like	0	0	15855	296	0
	Thioesterase/Thiol Ester Dehydrase-Isomerase	0	0	38098	226	0
	Xylose Isomerase-Like	0	0	18333	196	0
	TOTAL: 34	9	11	438108	4443	9

Select Task → | Download Network | Download Data Set

Sequence Similarity Networks

Download a Sequence Similarity Network of this superfamily (XGMML format).

Network downloads are XGMML files that are readable by program such as Cytoscape. In these networks, nodes represent proteins and edges represent pairwise similarities better than a given *E*-value cutoff. Additionally, these networks contain several attributes with data from the SFLD.

Select any restriction to apply to your network.

Maximum number of edges: 250K (•) 500K ( ) 750K ( ) all edges ( )

E-value Cutoff 1e-20

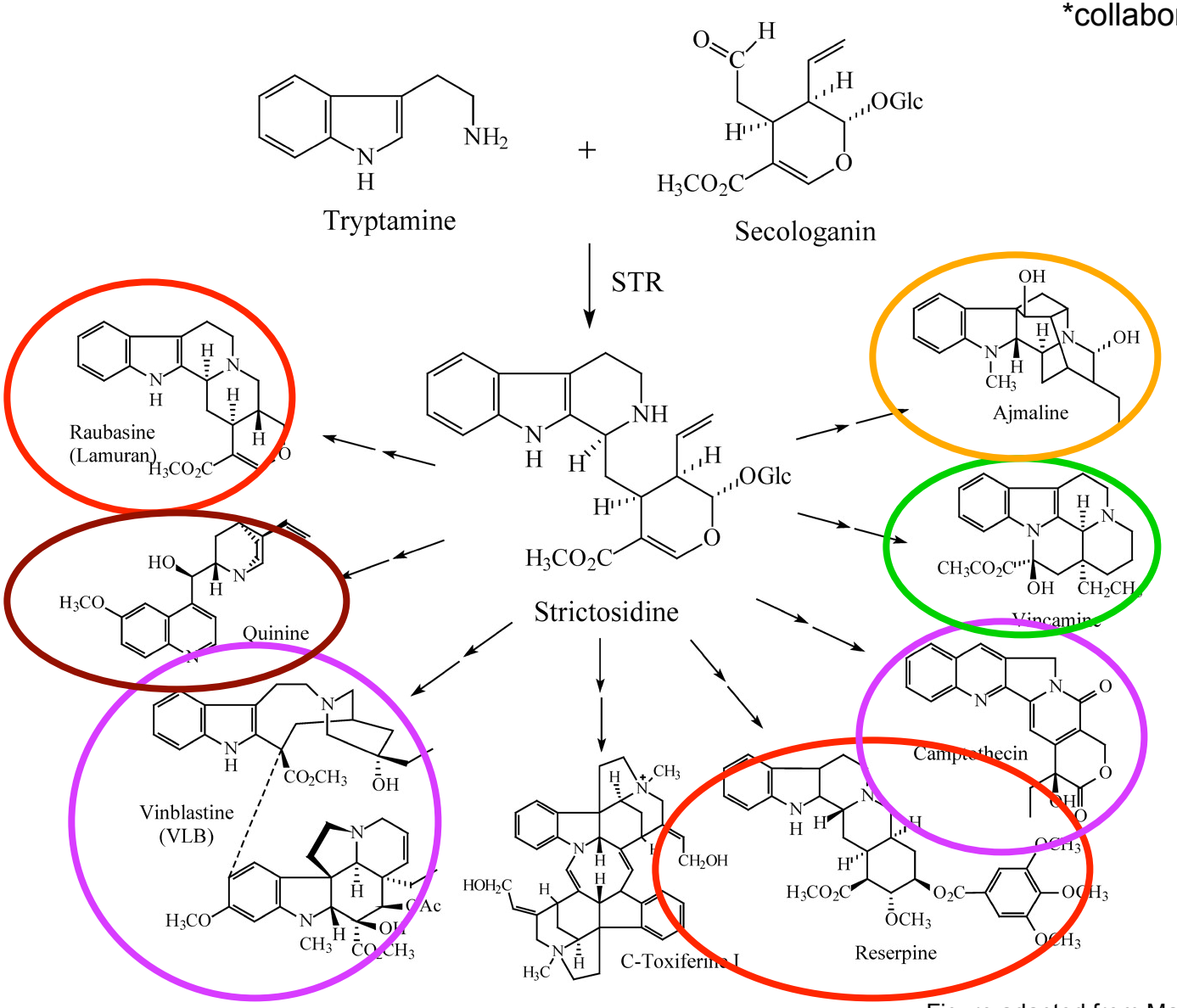
Full sequence similarity network is not available for this superfamily due to its size.

Repnet: 50% ID (6-Jun-2014)

What sequences to align:

The search for strictosidine variants from plants

*collaboration with Sarah O'connor



Cancer (vinblastine & topotecan

Malaria (quinine)

Hypertension (raubasine & reserpine)

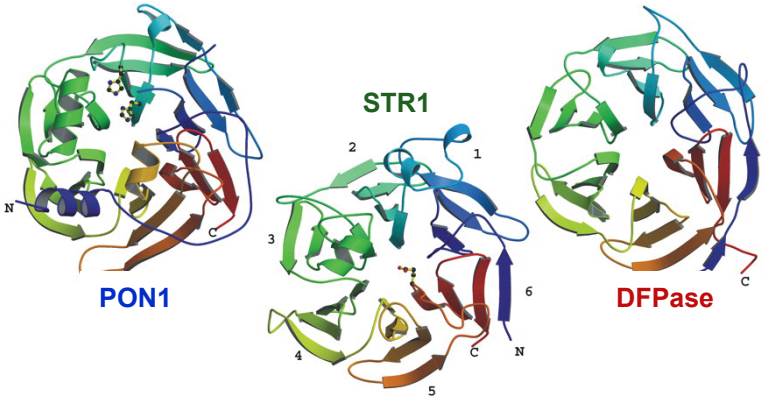
Schizophrenia (hi-dose reserpine)

Disturbed cerebral blood flow (vincamine)

Antiarrhythmia (ajmaline)

Figure adapted from Ma X et al, The Plant Cell, 18:907 (2006)

Nucleophilic attack 6-bladed β-propeller (N6P) SF



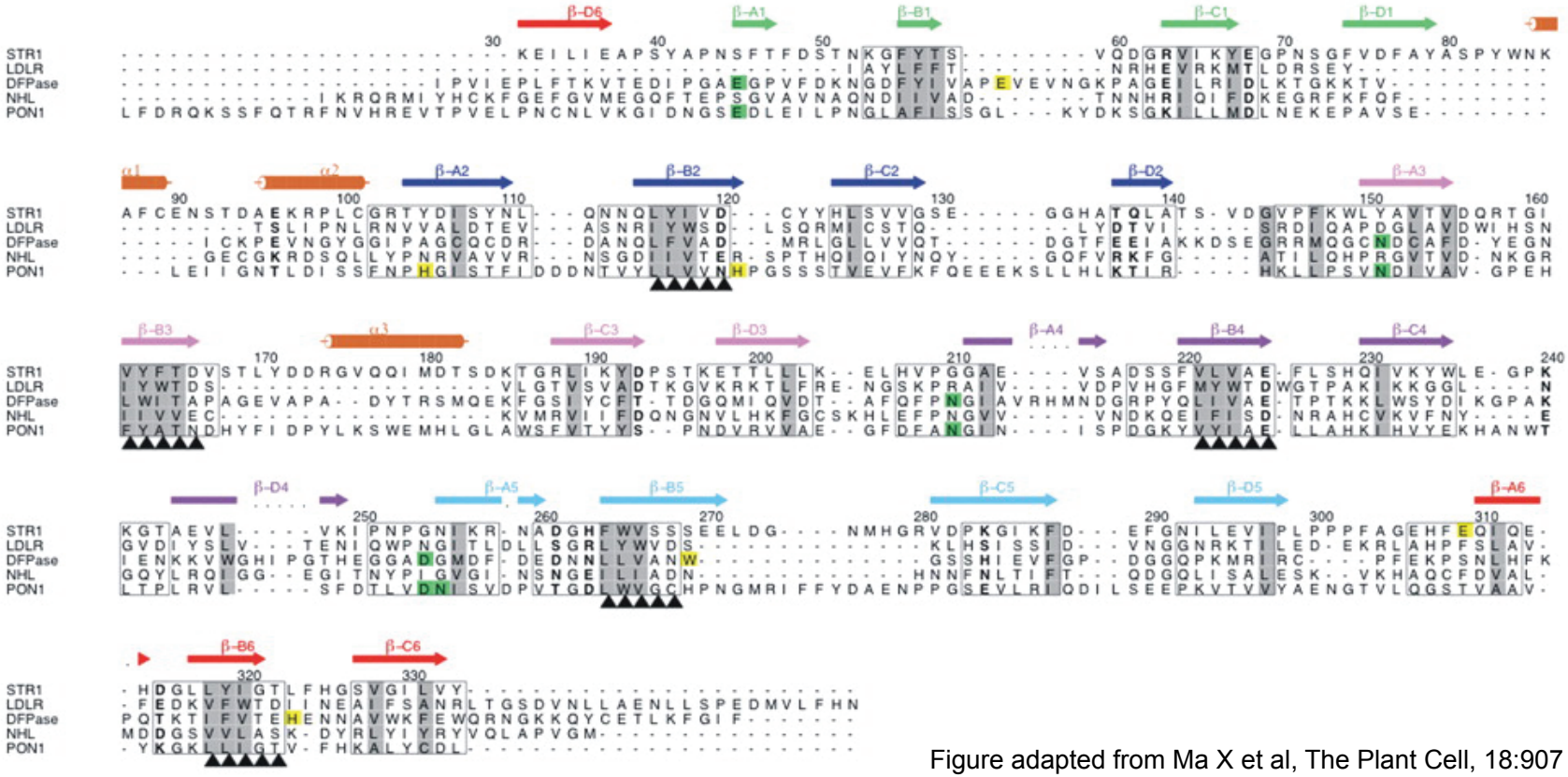
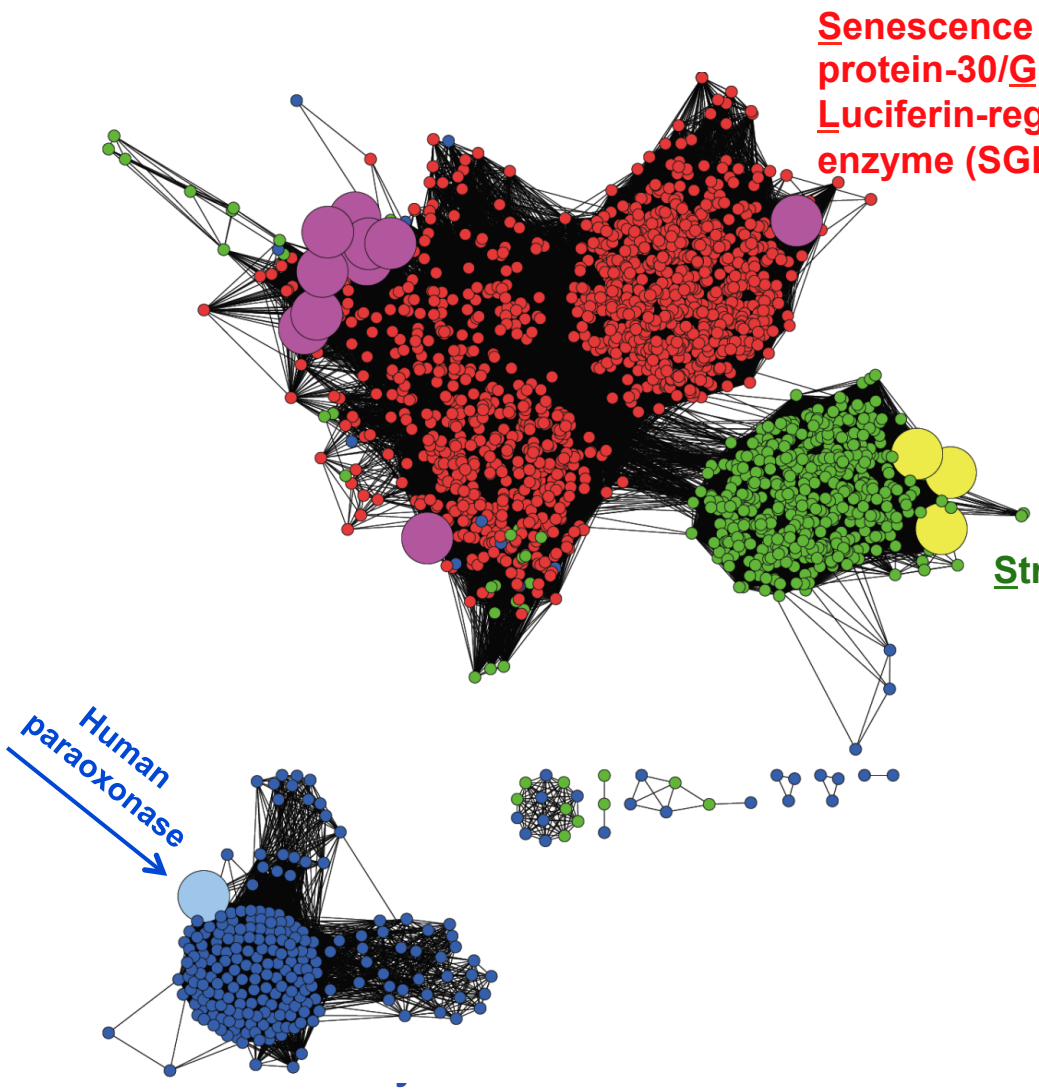


Figure adapted from Ma X et al, The Plant Cell, 18:907 (2006)

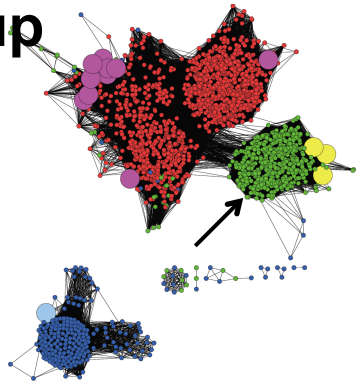
Characterized proteins are outliers in the N6P SF



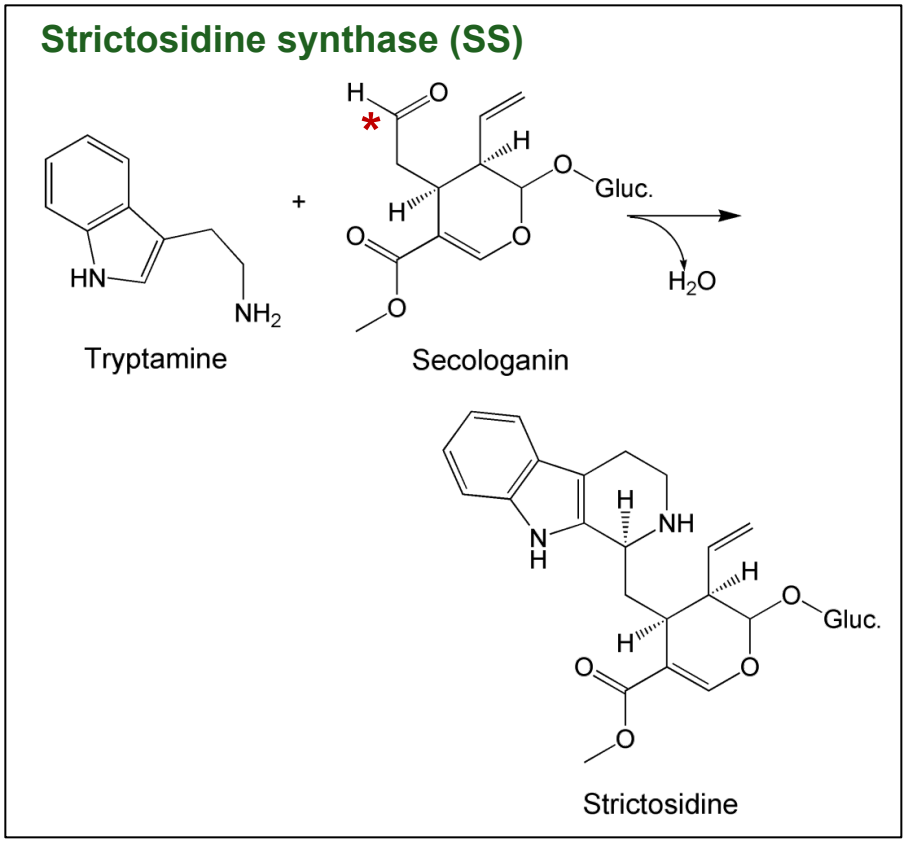
2,096 sequences
E-value threshold = 1 x 10^{-10}
Large nodes: experimentally characterized
Colored by PFAM model

Hicks et al, Proteins, 79:3082 (2011)

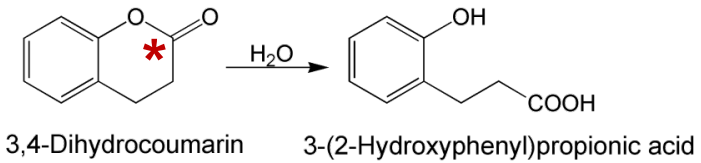
N6P catalyzes several different reactions involving nucleophilic attack on a double-bonded electrophilic group

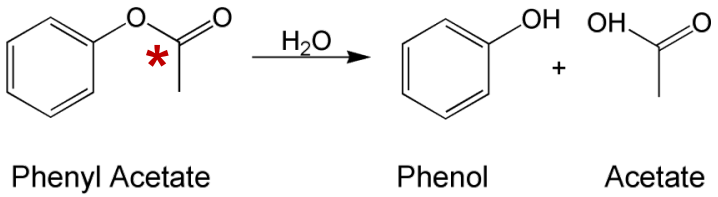


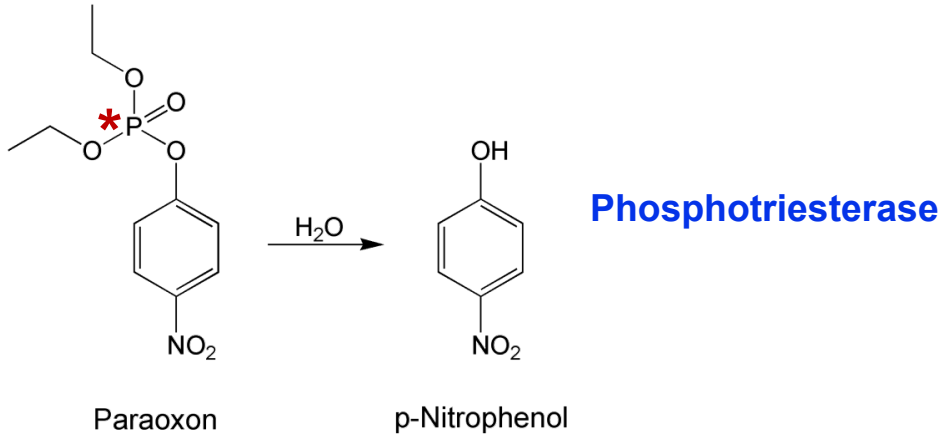
Strictosidine synthase (SS)

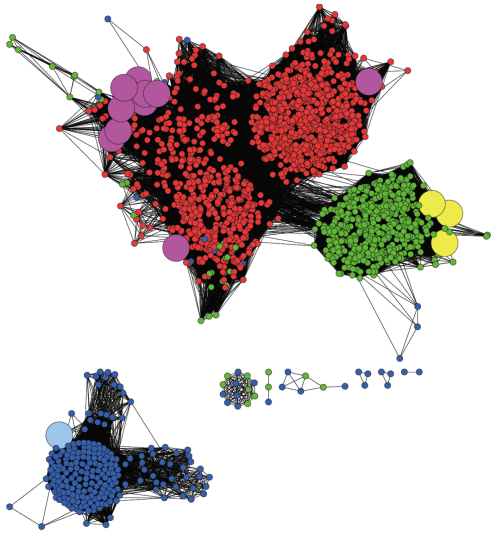


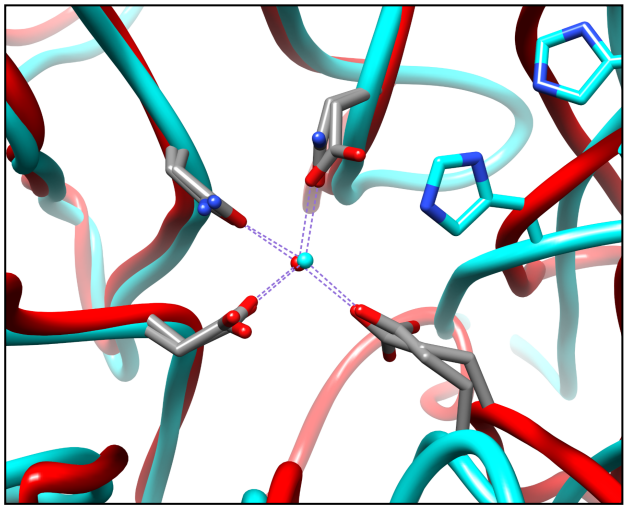
Characterized SS enzymes are functional outliers

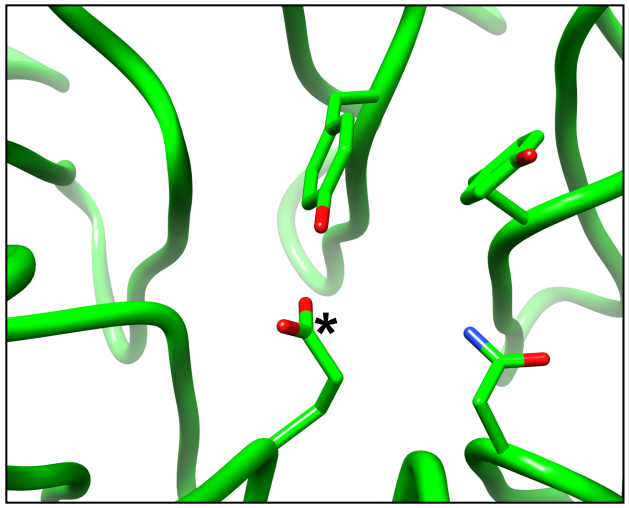






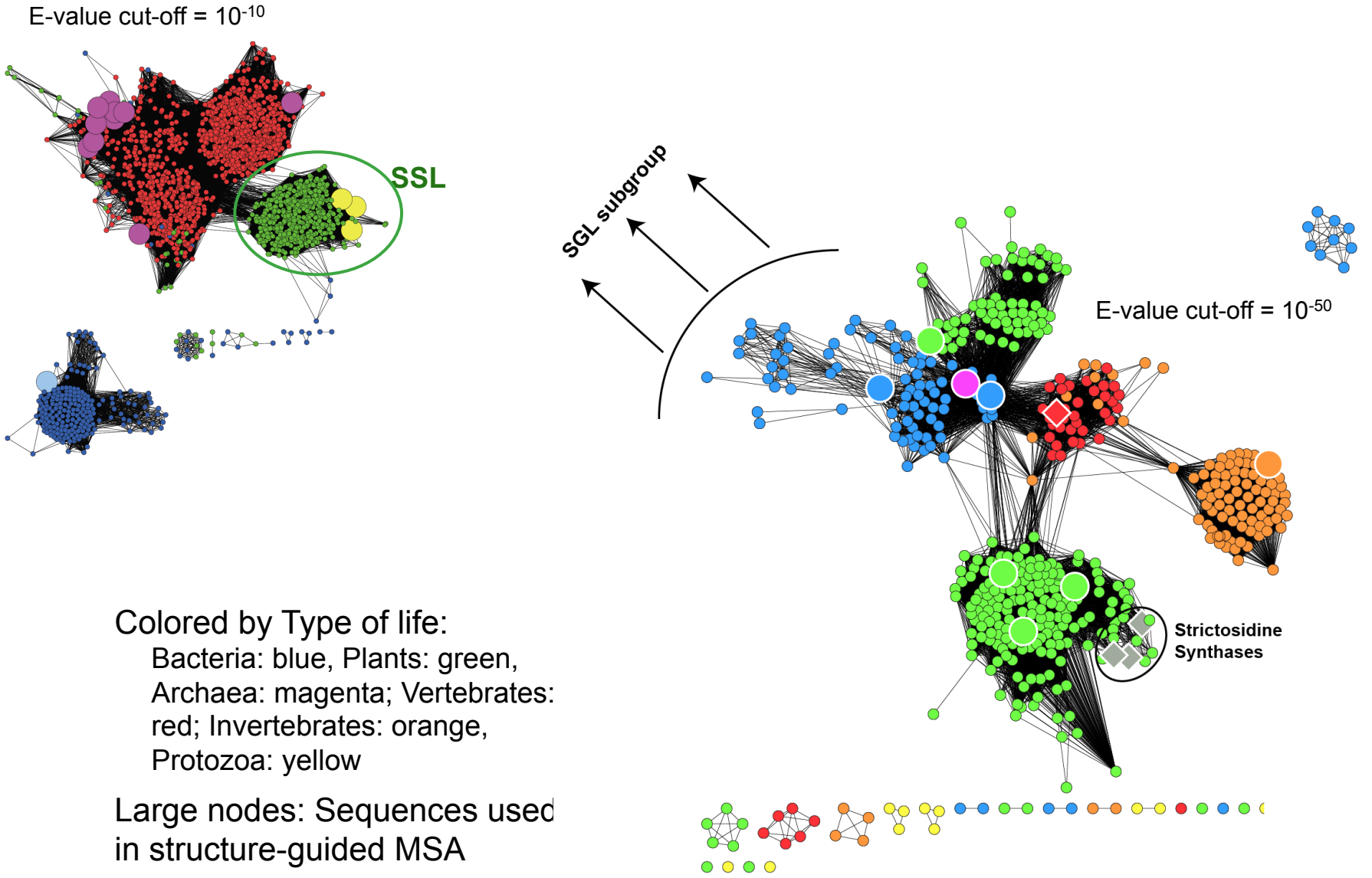






Active sites of characterized SS enzymes are outliers as well

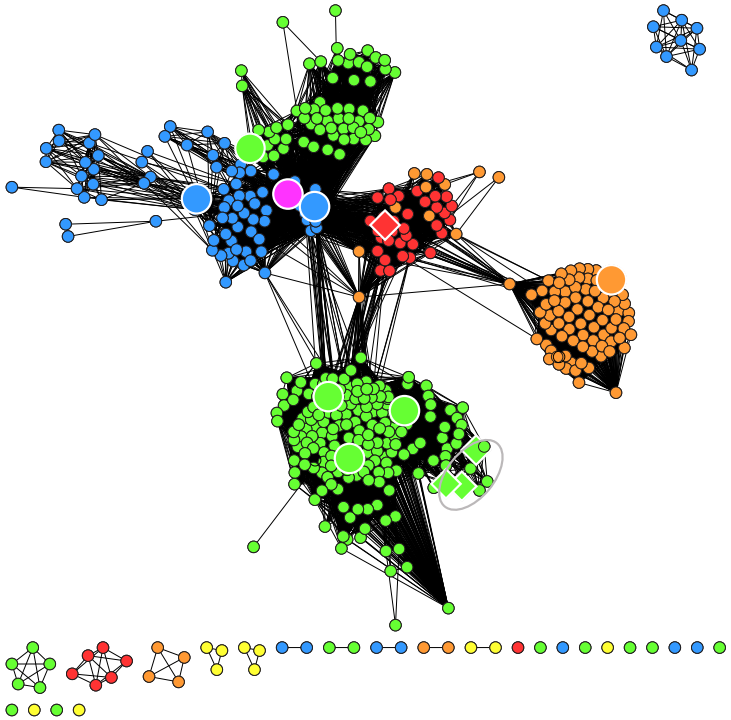
Many SSL proteins are not from plants



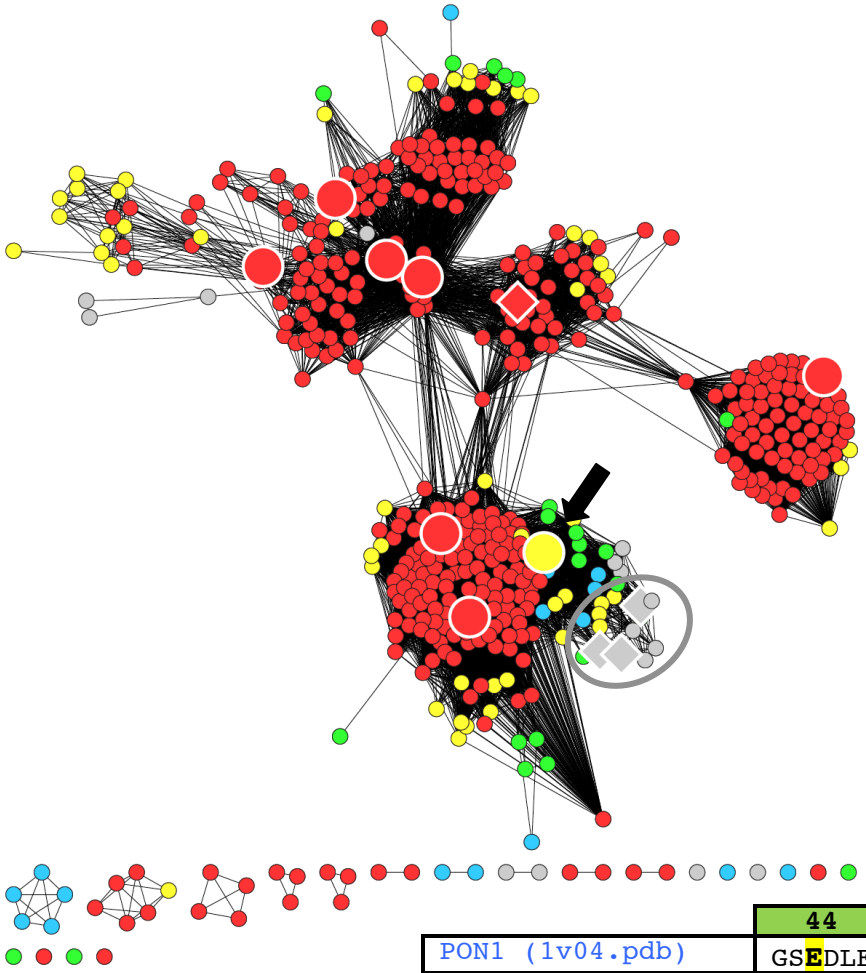
Colored by Type of life:

Bacteria: blue, Plants: green, Archaea: magenta; Vertebrates: red; Invertebrates: orange, Protozoa: yellow

Large nodes: Sequences used in structure-guided MSA



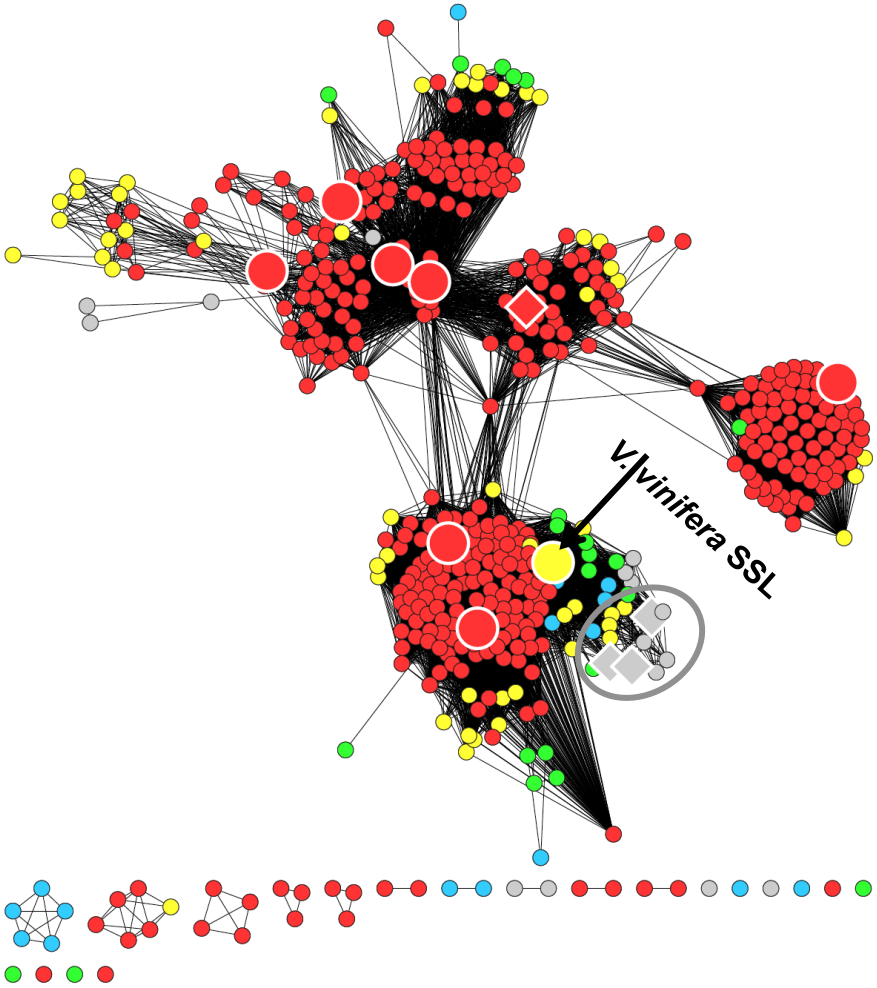
	44	151	210	254	309
PON1 (1v04.pdb)	GS**E**DLE	SV**N**DIVAVG	DVRVVAEGFDFA**N**GINISP	LV**D**NISVD	QGSTVAAV
Drp35 (2dg1.pdb)	QL**E**GLN	CI**D**DMVFDS	TVTPIIQNISVA**N**GIALST	GP**D**SCCID	LRSTHPQF
DFPase (1pjx.pdb)	GA**E**GPV	GC**N**DCAFDY	QMIQVDTAFQFP**N**GIAVRH	GA**D**GMDFD	EKPSNLHF
SS (2fpb.pdb)	**APNSFT**	**WLYAVTVDQ**	**ETTLLLKELHVPGGAEVSA**	**NPGNIKRN**	**EHFEQIQE**
gi|147772032	GP**E**AIA	FL**N**AVDVDQ	EVTVLLRGLGGAGGVTISK	TP**D**NIKRN	KTISEVQE
gi|22326950	GP**E**SVA	FT**N**DLDIAD	KAVVLVSNLQFP**N**GVSISR	HP**D**NVRTN	RSVSEVEE
gi|125556119	GP**E**SVA	FT**N**GVDIDQ	QVTVLQSNITYP**N**GVAISA	YP**D**NVRPD	RP-TEVMD
gi|24308201	GP**E**SIA	FV**N**DLTVTQ	EVKVLLDQLRFP**N**GVQLSP	FP**D**NIRPS	TYISEVHE
gi|1280434	GP**E**CLI	IF**N**GVTVSK	VSEVLLDELAFA**N**GLALSP	LP**D**NLTPD	T-ISHVLE
gi|125559158	AP**E**DVY	FA**D**AAIEAS	EASVVLDGLGFA**N**GVALPP	NP**D**NIRLG	NMVTSVTE
gi|111017930	GP**E**DVA	AC**N**NSAVGR	ETDLLAEGLQFA**N**GVGLAS	IP**D**NMTSQ	P-VTGVRE
gi|15596490	GP**E**DTA	FT**D**DLDIAS	KTEVLLKDLYFA**N**GVALSA	LP**D**NLQGD	RMITSAKP
gi|52549517	GP**E**DVA	LT**D**DVDIAA	TTRLVLNNLYFA**N**GVAVSP	FP**D**GISSN	Q-ITSVQE



E-value cut-off = 10^{-50}
516 SSL subgroup sequences
Median alignment length = 297 residues; median percent identity = 41%
Colored by # of metal coordinating residues
Red = 4
Yellow = 3
Green = 2
Cyan = 1
Gray = 0

	44	151	210	254	309
PON1 (1v04.pdb)	GS**E**DLE	SV**N**DIVAVG	DVRVVAEGFDFA**N**GINISP	LV**D**NISVD	QGSTVAAV
Drp35 (2dg1.pdb)	QL**E**GLN	CI**D**DMVFDS	TVTPIIQNISVA**N**GIALST	GP**D**SCCID	LRSTHPQF
DFPase (1pjx.pdb)	GA**E**GPV	GC**N**DCAFDY	QMIQVDTAFQFP**N**GIAVRH	GA**D**GMDFD	EKPSNLHF
SS (2fpb.pdb)	**APNSFT**	**WLYAVTVDQ**	**ETTLLLKELHVPGGAEVSA**	**NPGNIKRN**	**EHFEQIQE**
gi|147772032	GP**E**AIA	FL**N**AVDVDQ	EVTVLLRGLGGAGGVTISK	TP**D**NIKRN	KTISEVQE
gi|22326950	GP**E**SVA	FT**N**DLDIAD	KAVVLVSNLQFP**N**GVSISR	HP**D**NVRTN	RSVSEVEE
gi|125556119	GP**E**SVA	FT**N**GVDIDQ	QVTVLQSNITYP**N**GVAISA	YP**D**NVRPD	RP-TEVMD
gi|24308201	GP**E**SIA	FV**N**DLTVTQ	EVKVLLDQLRFP**N**GVQLSP	FP**D**NIRPS	TYISEVHE
gi|1280434	GP**E**CLI	IF**N**GVTVSK	VSEVLLDELAFA**N**GLALSP	LP**D**NLTPD	T-ISHVLE
gi|125559158	AP**E**DVY	FA**D**AAIEAS	EASVVLDGLGFA**N**GVALPP	NP**D**NIRLG	NMVTSVTE
gi|111017930	GP**E**DVA	AC**N**NSAVGR	ETDLLAEGLQFA**N**GVGLAS	IP**D**NMTSQ	P-VTGVRE
gi|15596490	GP**E**DTA	FT**D**DLDIAS	KTEVLLKDLYFA**N**GVALSA	LP**D**NLQGD	RMITSAKP
gi|52549517	GP**E**DVA	LT**D**DVDIAA	TTRLVLNNLYFA**N**GVAVSP	FP**D**GISSN	Q-ITSVQE

The large-scale context suggests the great majority of SSL proteins are not SSs



Vitus vinifera SSL lacks SS activity but shows low levels of hydrolase activity typical of the SSL and arylesterase subgroups

Challenges

Technical issues

- > Extreme sequence diversity
- > Complex insert patterns and multiple domain architectures within a SF

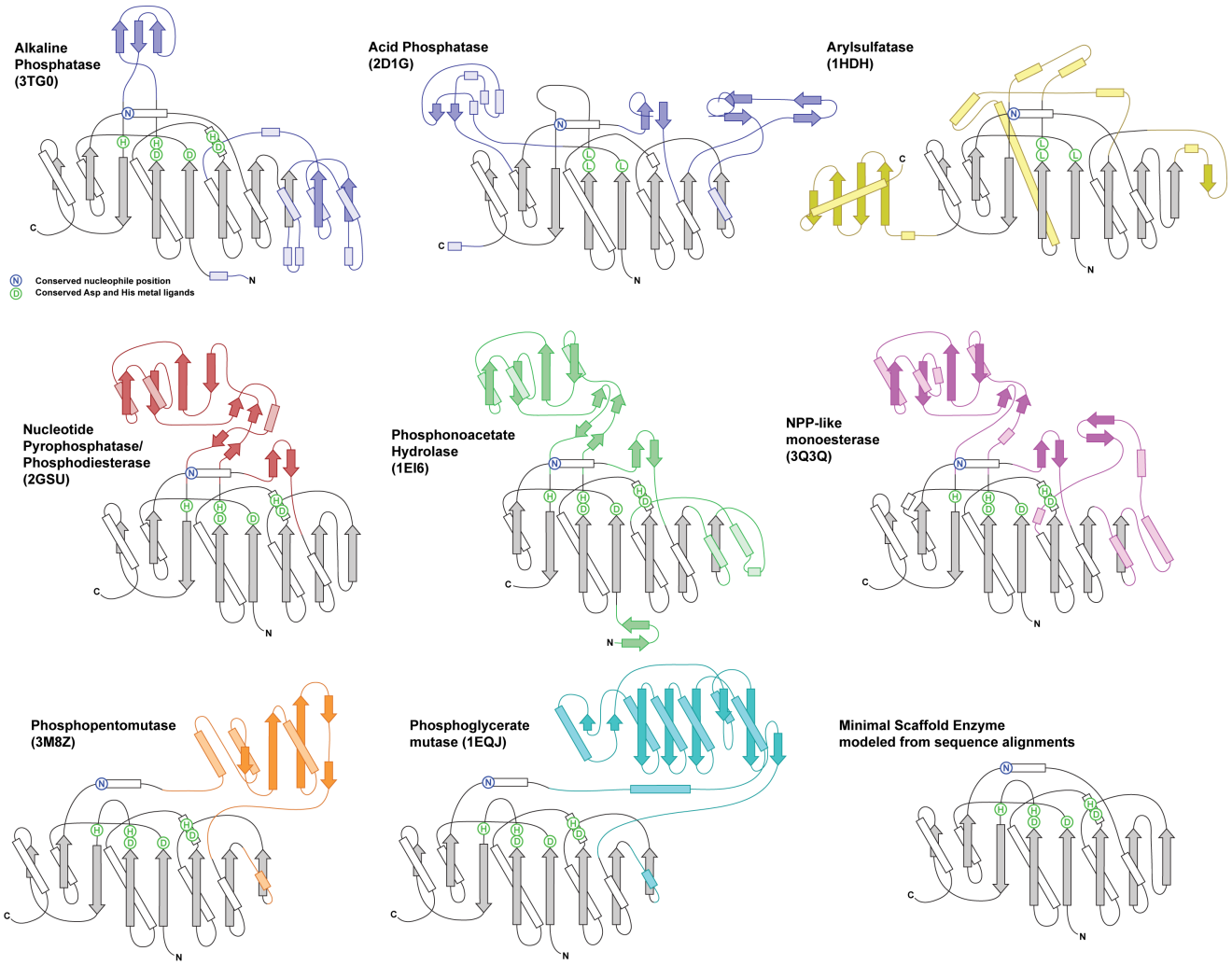
Experimentally characterized proteins only poorly sample the available sequence and structure space

Especially for divergent proteins of functionally diverse enzyme SFs, clustering by sequence and structure similarity fails to track well with functional boundaries

Uniqueness of structure-function relationships in individual SFs complicates development of general solutions

Alkaline phosphatase superfamily: Inserts to the common core distinguish known reaction classes

40,000 sequences



Many of these inserts are unrelated

> Locations vary

> Multiple insertions in a single subgroup

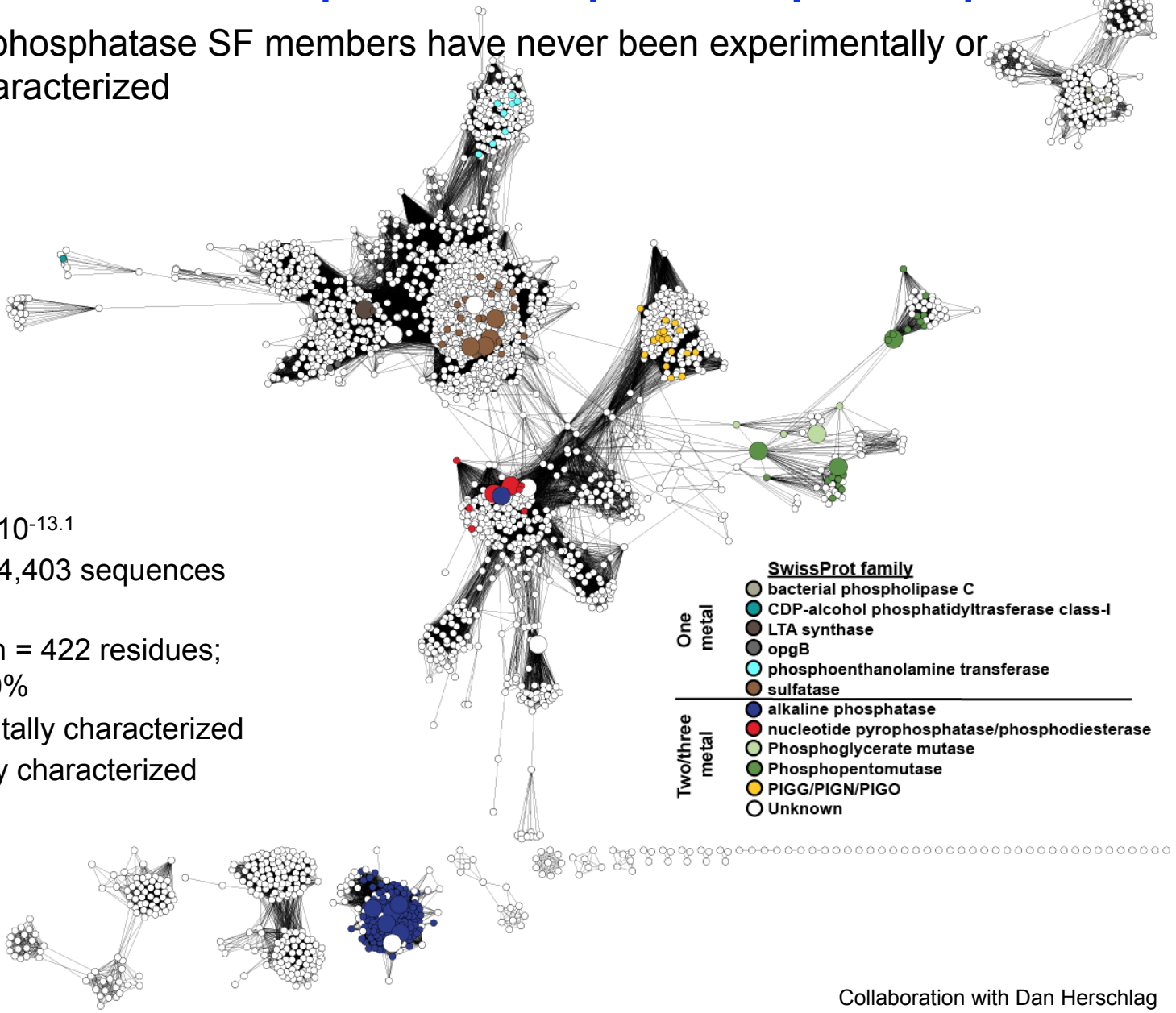
> Structural insert patterns fail to track with variations in reaction or mechanism

*Collaboration with Dan Herschlag

What sequences to align?

How well do characterized proteins sample SF sequence space?

Most alkaline phosphatase SF members have never been experimentally or structurally characterized



E-value threshold = 1 x $10^{-13.1}$

4590 nodes represent 14,403 sequences (40% ID filtered)

Median alignment length = 422 residues;
Median pairwise ID = 30%

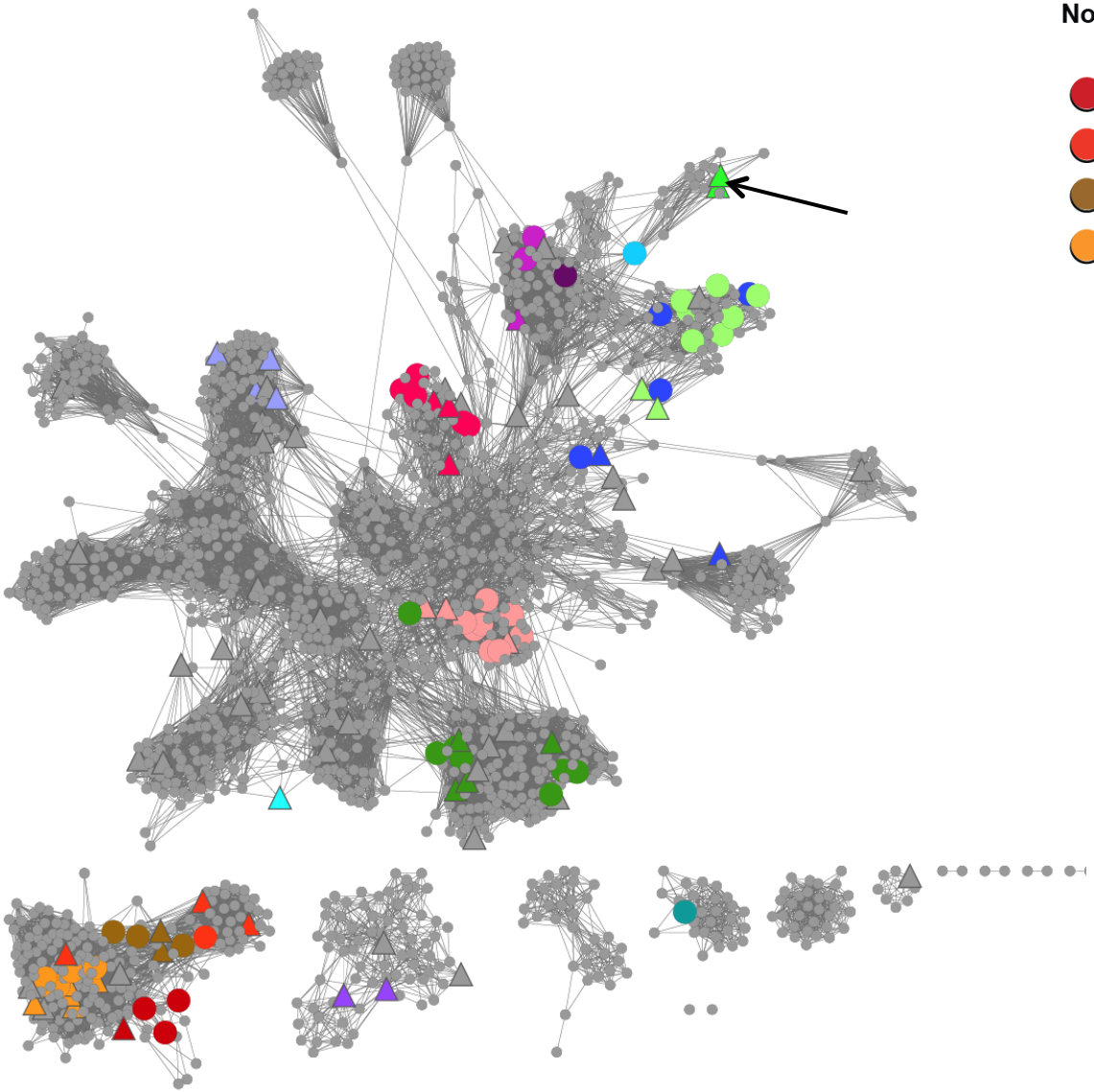
Small nodes: experimentally characterized

Large nodes: structurally characterized

Collaboration with Dan Herschlag

Even very well-studied SFs are minimally characterized

< 2% of GSTs experimentally shown to catalyze GST-like reactions



Mashiyama et al, PLoS Biol, 12: :e1001843 (2014)

Node color: Swiss-Prot family annotation

- Alpha
- Beta
- Omega
- Mu
- DHAR
- Phi
- Pi
- FAX
- Tau
- Sigma
- HSP26
- Theta
- Lambda
- Zeta
- chloride channel CLIC

"New" classes

- Nu
- Xi

cytosolic Glutathione transferase (GST) SF

E-value threshold = $1x10^{-13}$

1,568 nodes represent 13,000 sequences (50% ID filtered)

Each node contains 1-930 sequences

Median edge *E*-value = $4x10^{-25}$

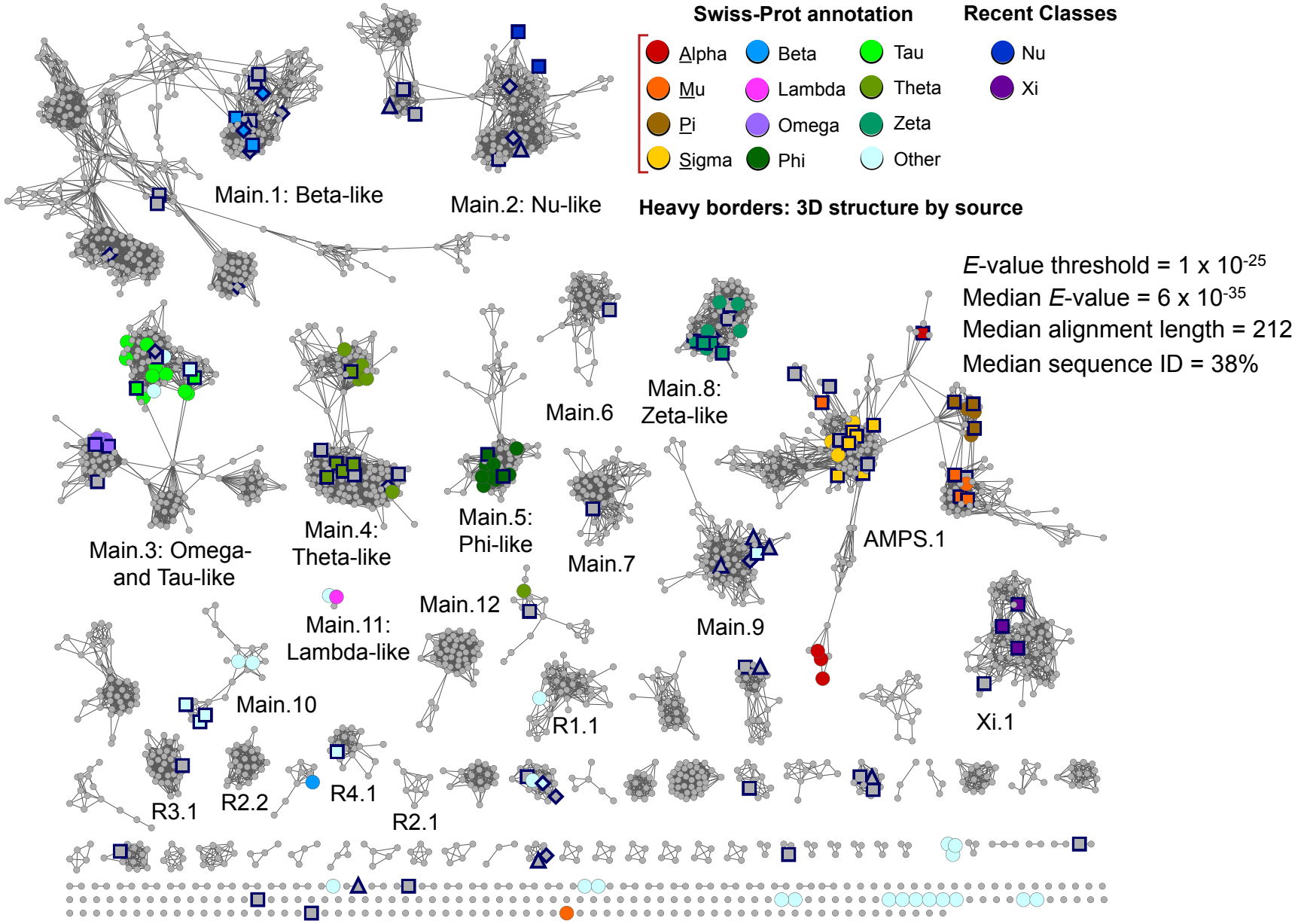
Median alignment length = 210 residues

Large nodes colored by SwissProt classification if >50% of sequences in each node belong to that class

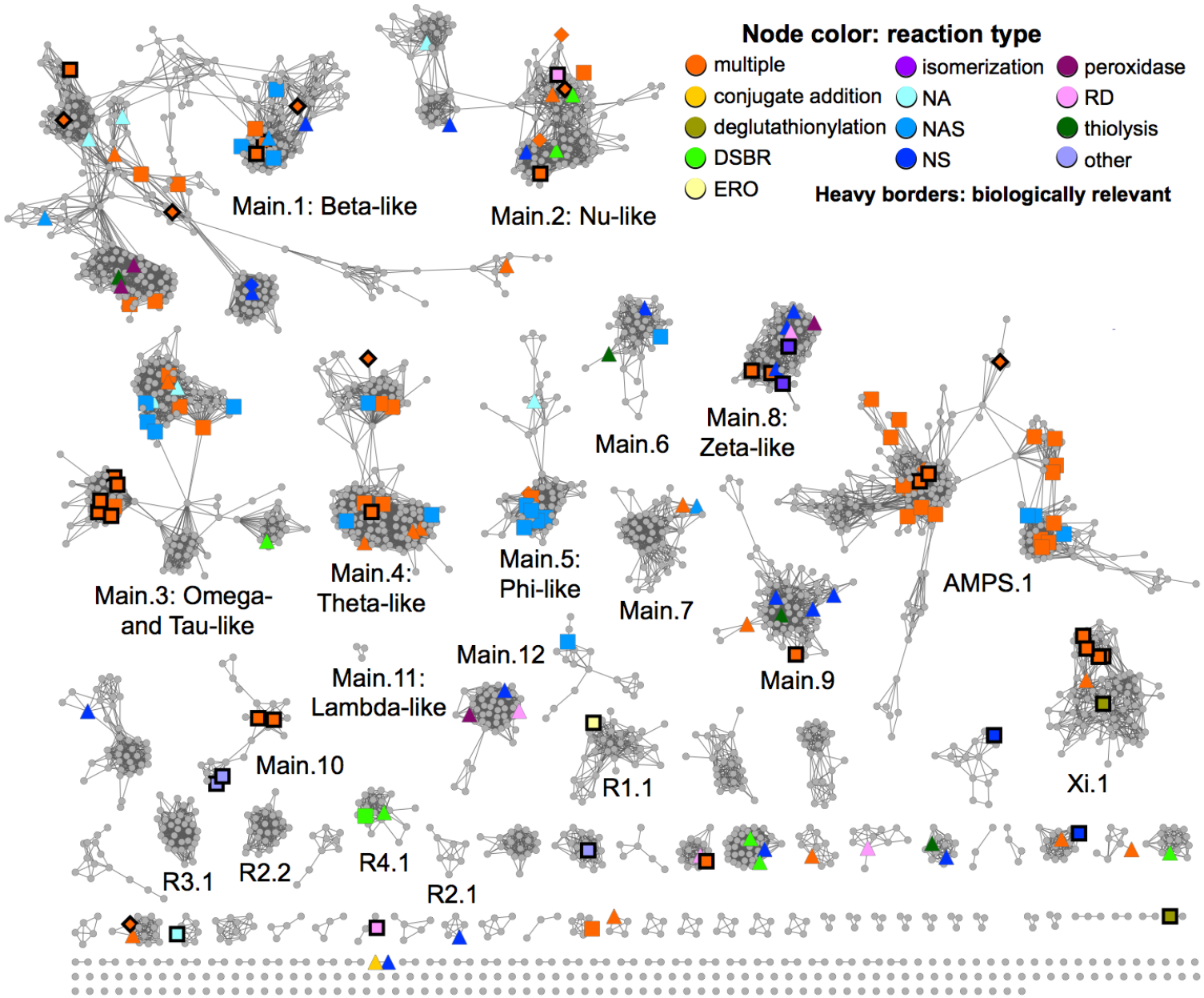
Gray: Not classified

Triangles: PDB structure

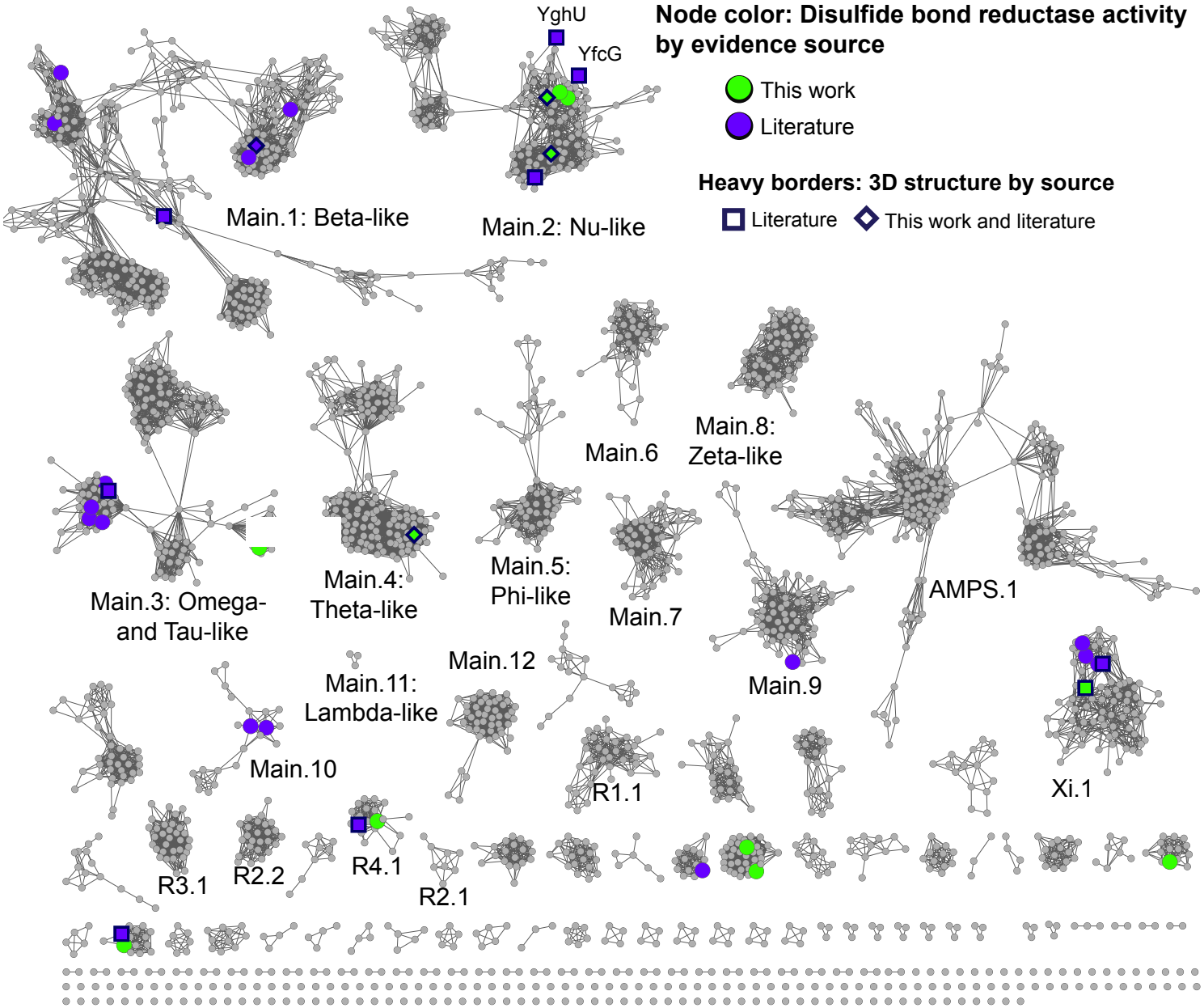
Many new classes yet to be discovered



Reaction types fail to track with similarity clusters

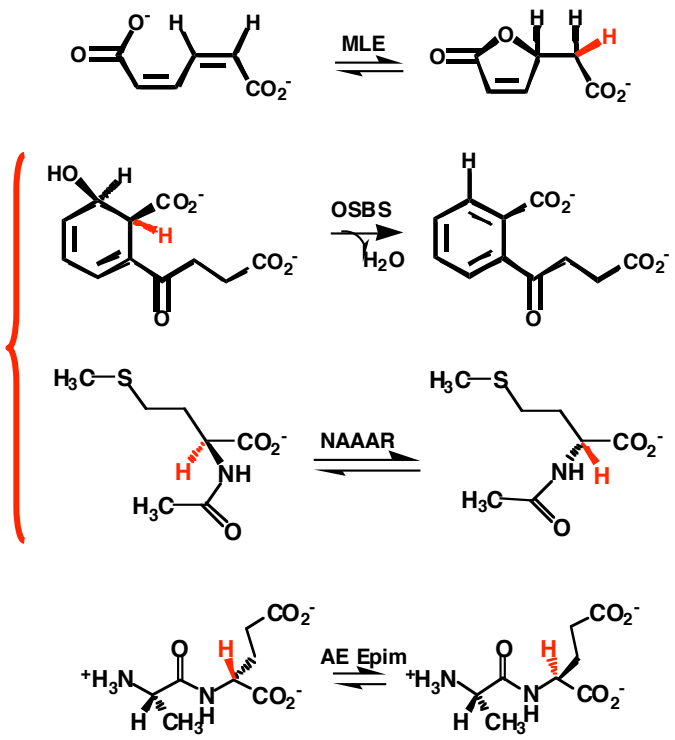


Promiscuity likely more widespread than we know

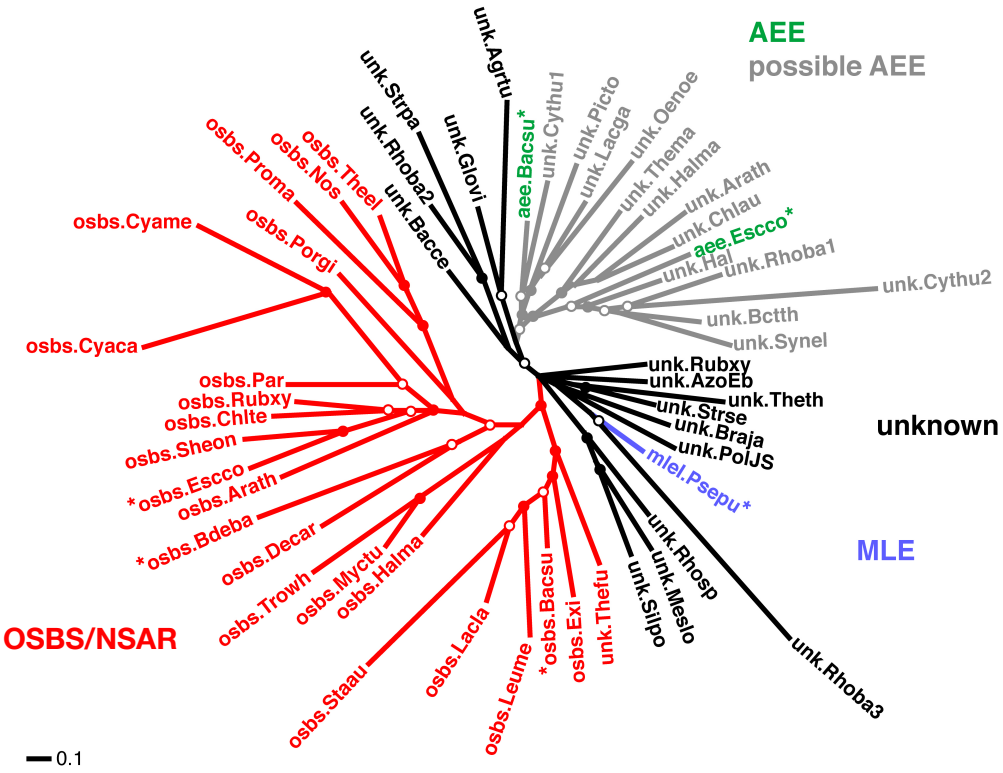


Reaction families within a SF evolve at different rates

Enolase SF



Some OSBS' are promiscuous



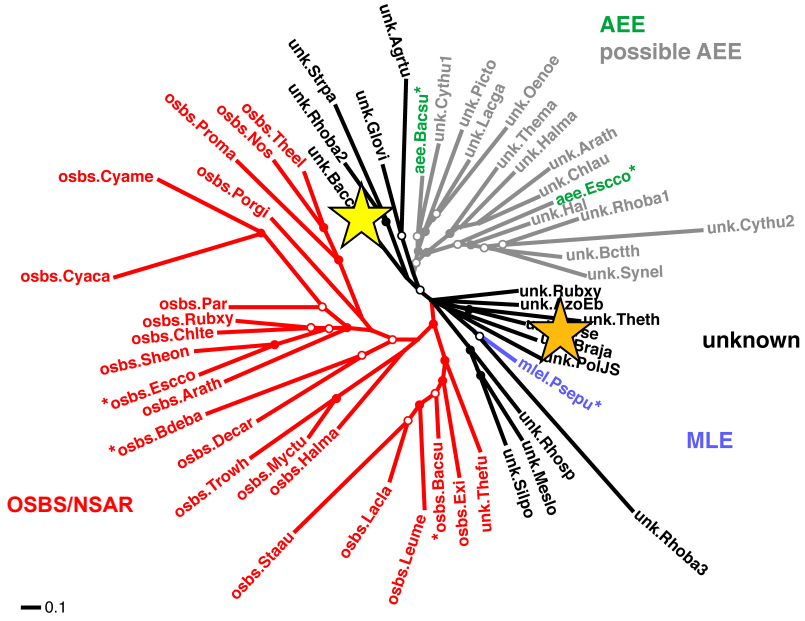
Alignment: sequences of 505 orthologs
filtered to 40% ID (54 seqs)

Bayesian tree (Mr. Bayes)
branch confidence:
- ● ≥ 0.95
- ○ 0.7-0.94

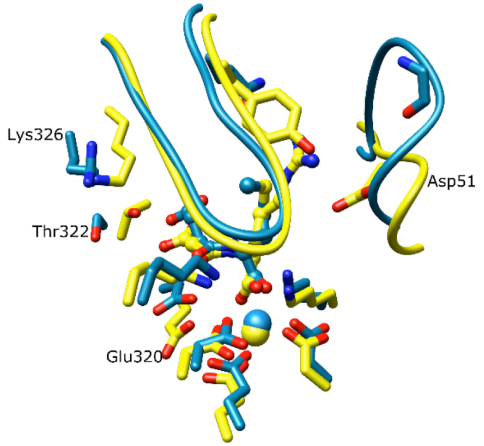
*structurally characterized

Glasner et al, JMB 360: 228 (2006)

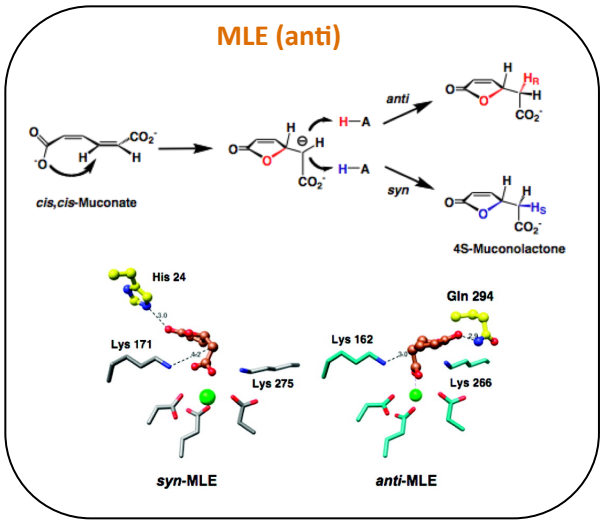
“Pseudo-convergent” evolution of the same reaction from different intermediate ancestors in the SF tree



Dipeptide epimerase
MLE (syn)
OSBS/NSAR-1
NSAR-2
Unknown



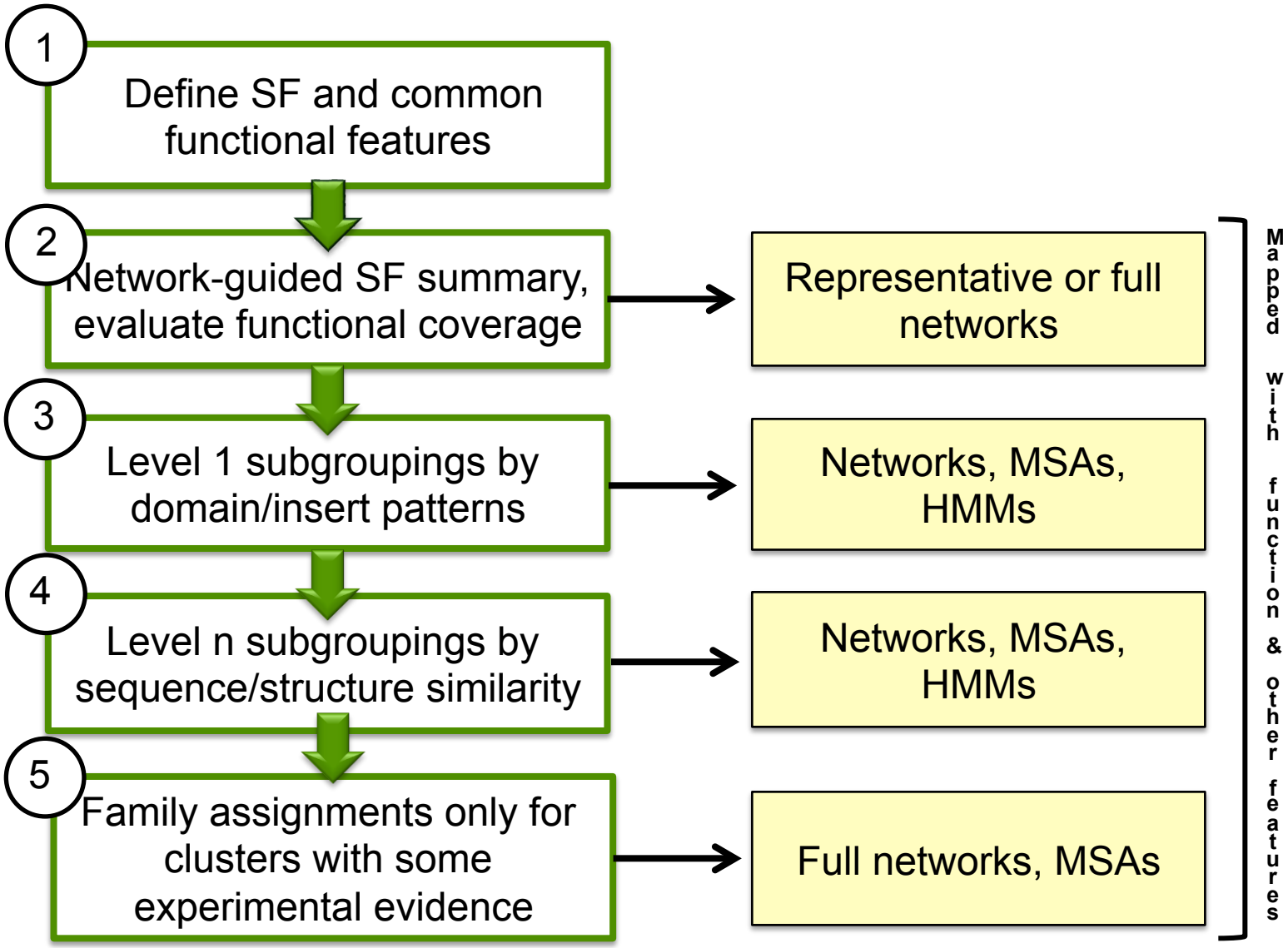
Song et al, Nat Chem Biol, 3:486 (2007)



Sakai et al, Biochem, 48:2569 (2009)

What sequences to align

SFLD SF curation guided by sequence, structure, and functional information



Final thoughts

Alignment methods with more sophisticated tools for choosing "representative" sequences to include in an MSA would contribute significantly to these difficult problems

> Especially important for both manual and automated curation communities

Automated function prediction experiments (CAFA) suggest the value of methods that allow for incorporation of functionally relevant features

> Already included in some phylogenomic methods

Similarity network methods could supplement current MSA approaches but need more rigorous development

We can supply manually curated Gold Standard sets for evaluation of new methods

Acknowledgments

Eyal Akiva, PhD
Holly Atkinson, PhD*
Alan Barber, PhD*
Shoshana Brown, PhD
Gemma Holliday, PhD
Michael Hicks, PhD*
Florian Lauck*
Susan Mashiyama, PhD*
Elaine Meng, PhD
David Mischel
Rebecca Davidson
Alexandra Schnoes, PhD*
Doug Stryke*
Jack Yu
Jeff Yunes

*Former members

Collaborators

USCF Resource for Biocomputing, Visualization & Informatics
Tom Ferrin
John Morris

N6P Superfamily
Sarah O'Connor (Danforth Center)

Enolase Superfamily
John Gerlt (U of IL)
Matt Jacobson (UCSF)

CytGST Superfamily
Enzyme Function Initiative
Steven Almo (Einstein)
Richard Armstrong (Vanderbilt)

Alkaline phosphatase superfamily
Dan Herschlag (Stanford)
Jonathan Lassila

$$ NIH, NSF $$