University of California
San Francisco

Unraveling new protein functions and pathways using sequence similarity networks

Eyal Akiva

The intricacy of function prediction

- **The question challenge**

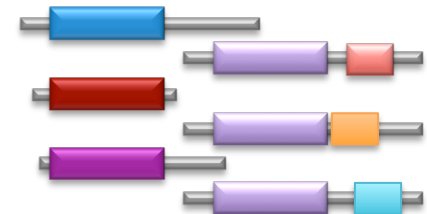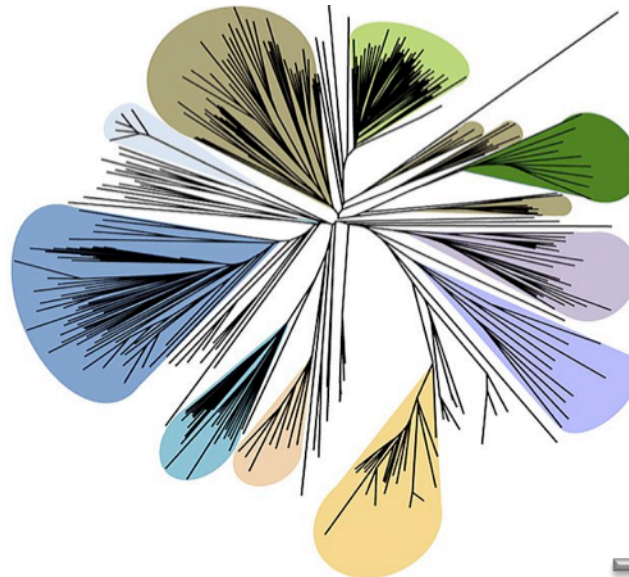- **Sequence – function correlation**
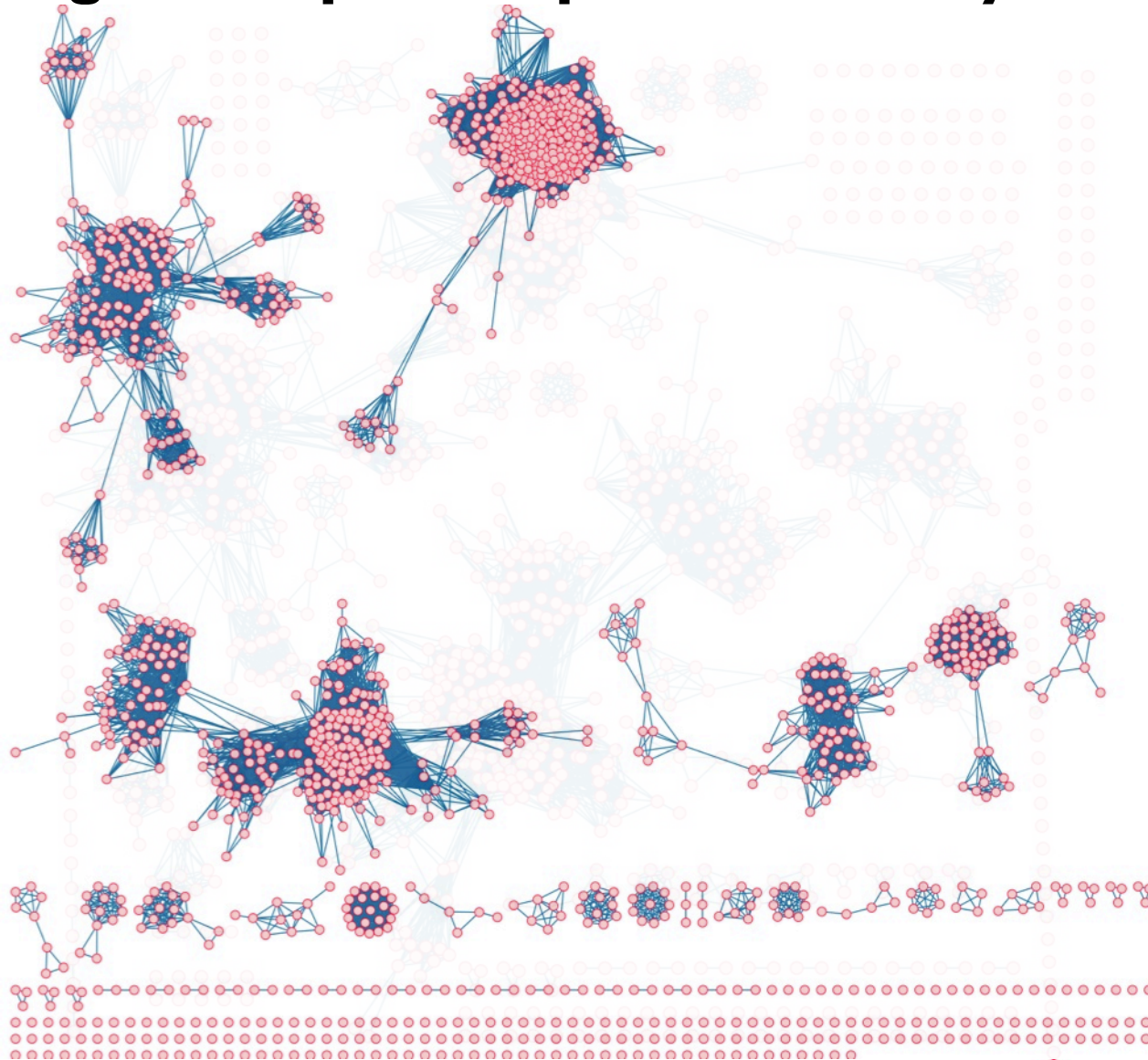
- **The platform challenge**

Classification:
A gateway for predicting function and enzymatic pathways

Mapping the sequence space

Mapping the sequence space: Similarity networks

Node   A protein   **Edge**   Pair-wise BLAST-based sequence similarity

Smoot ME et al, Bioinformatics. ;27(3):431 (2011)

Atkinson HJ, Morris JH, Ferrin TE, Babbitt PC. PLoS One. 4(2):e4345 (2009)

Representative protein similarity networks

proteins per node

- •   x=1
- •   2≤x≤5
- ●   6≤x≤20
- ●   21≤x≤100
- ●   x>100

PYTHOSCAPE

Node   **proteinS**   **Edge**   Pair-wise BLAST-based sequence similarity

Barber & Babbitt, Bioinf, 28:2845 (2012)
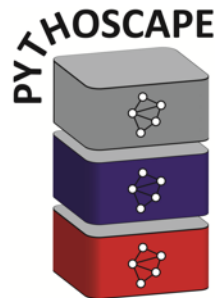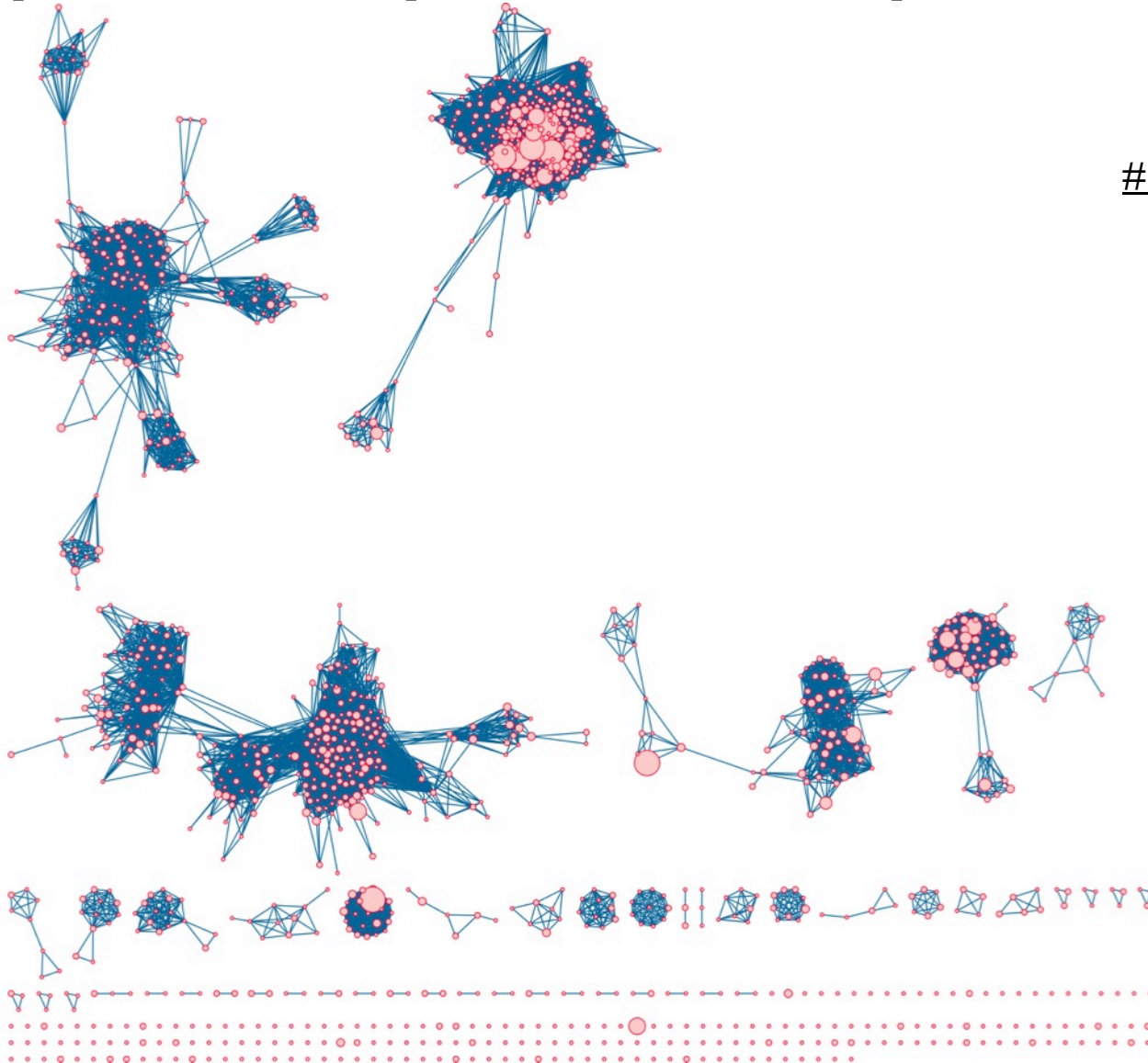
Similarity thresholds
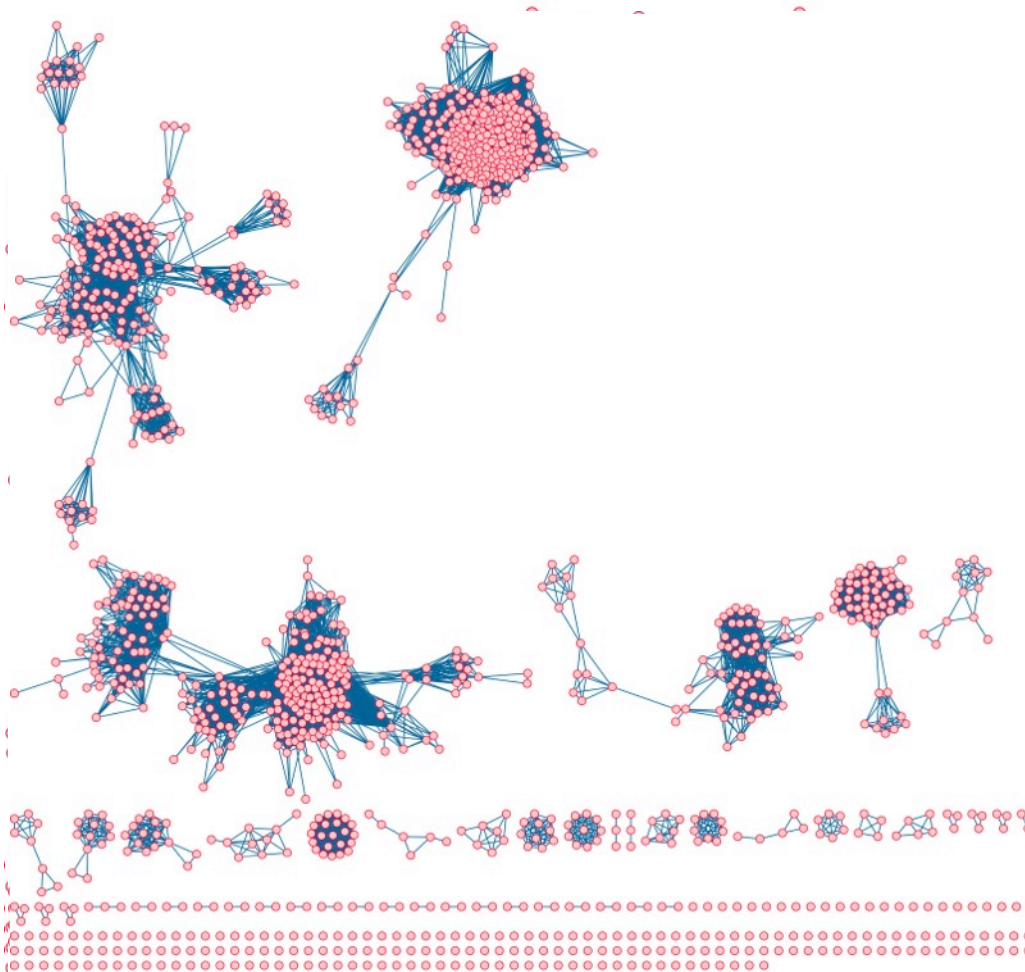
BLAST e-value

→ 1e-50

→ 1e-60

→ 1e-70

→ 1e-80

→ 1e-90

→ 1e-100

→ 1e-110

→ 1e-125

protein classification

Superfamily   A set of underlined evolutionarily related enzymes with a shared chemical function that maps to a conserved set of active site features

Evolutionary relations

shared reaction step

conserved set of active site features

Application I :

Function prediction

E-value threshold: 1e−44

- 🟥 D-arabinonate dehydratase
- 🟦 D-tartrate dehydratase
- 🟩 L-fuconate dehydratase
- 🟪 L-talarate/galactarate dehydratase
- 🟧 galactonate dehydratase
- 🟨 gluconate dehydratase
- 🟪 mandelate racemase
- ⬜ rhamnonate dehydratase
- 🟫 gut metagenome
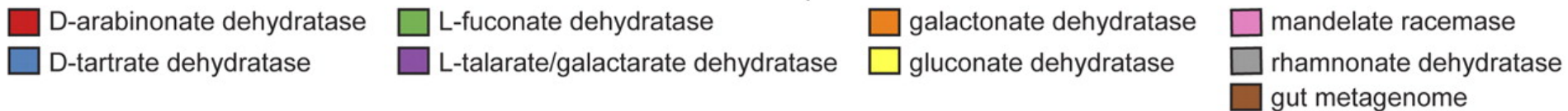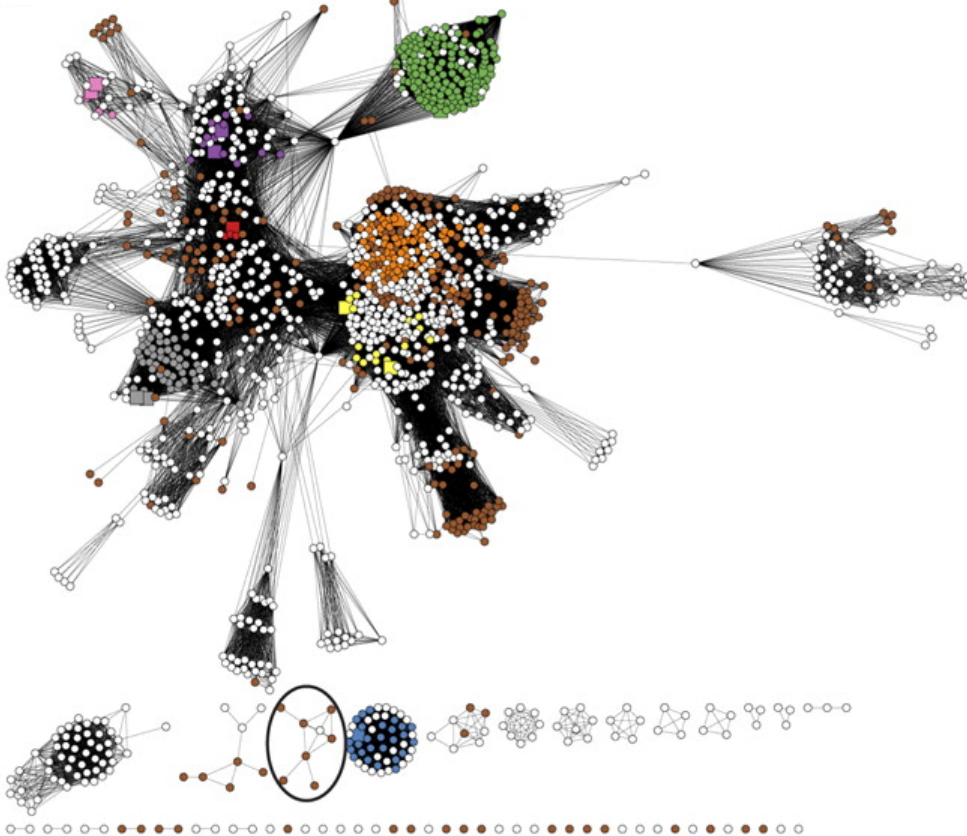
Sequence similarity network of acid-sugar dehydratases known or predicted to belong to enolase superfamily and human gut microbiome.

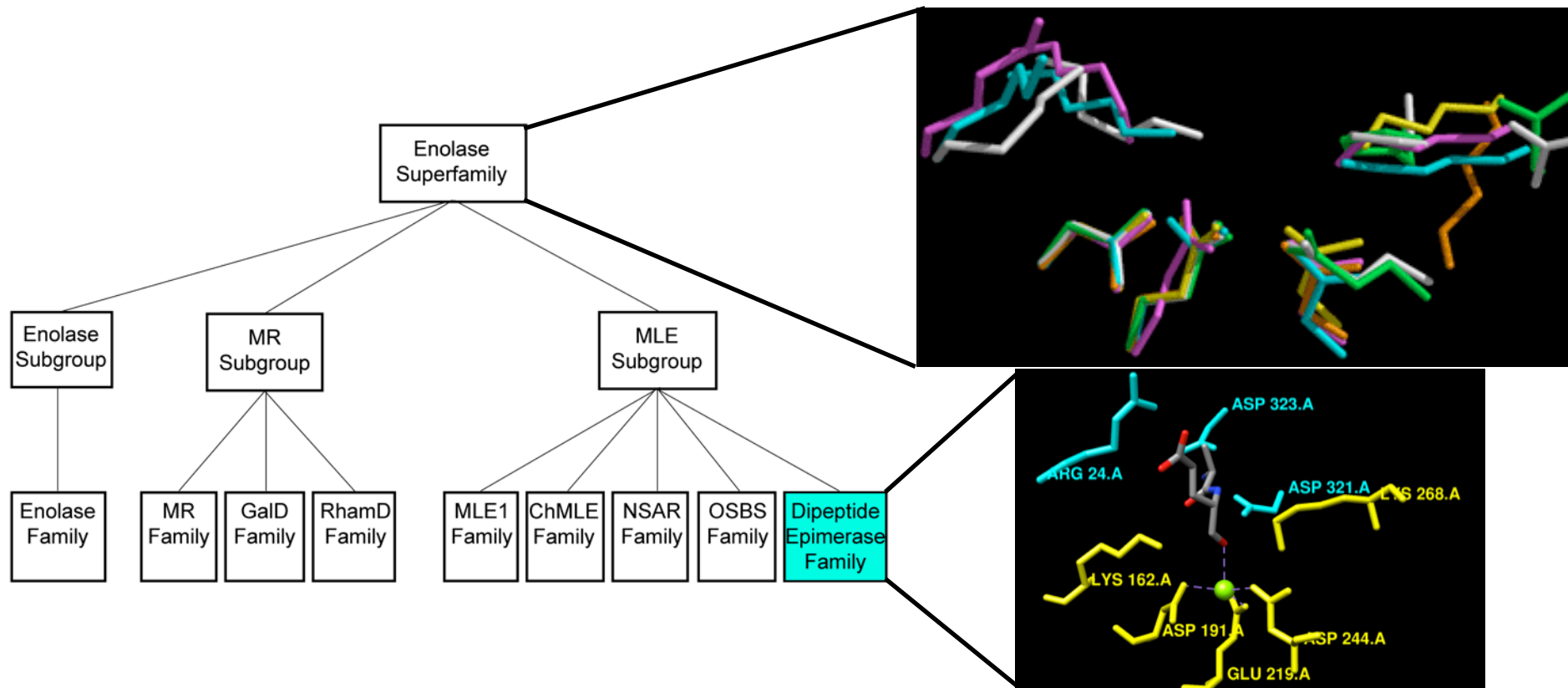Brown S D , and Babbitt P C J. Biol. Chem. 2012;287:35-42

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

A database that offers hierarchical classification of enzymes that relates specific sequence-structure features to specific chemical capabilities

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

A database that offers hierarchical classification of enzymes that relates specific sequence-structure features to specific chemical capabilities

Sample superfamilies

- Amidohydrolase
- Crotonase
- Enolase
- Glutathion-S-transferase
- Haloacid dehalogenase
- Isoprenoid Synthase Type I
- Peroxiredoxin
- Radical SAM

Query sequence
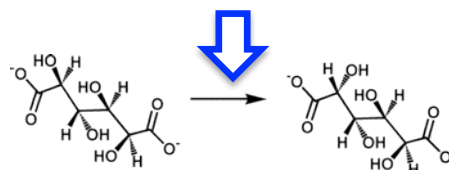
Position	Type	Function
196	His (H)	stabilizes transition state
240	Asp (D)	metal binding ligand
277	Glu (E)	metal binding ligand
311	Asp (D)	metal binding ligand
335	Lys (K)	abstracts alpha proton (base)

Main goals

Functional assignments of unknown enzymes and the elucidation of new pathways

JA Gerlt et al. *Biochemistry* **2011,** 50, 9950-9962
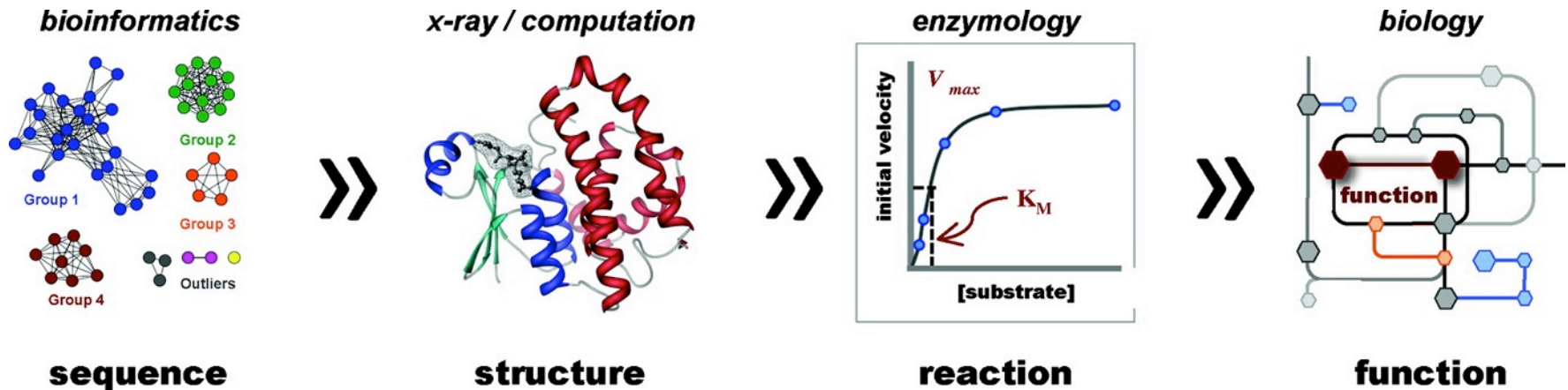
Summary

- Protein similarity networks are an attractive adjunct approach to multiple sequence alignments and phylogenetic trees.

- Application to classification, function prediction, target selection and elucidating new pathways.

- Serves regularly as a platform for both protein classification in the **SFLD** database as a a pivotal bioinformatic tool for the **EFI** efforts

Summary

- **The platform challenge**

- **Sequence – function correlation**

- **The question challenge**

Acknowledgements