Genomic Enzymology: Discovery of Novel Enzymes in Novel Pathways

John A. Gerlt

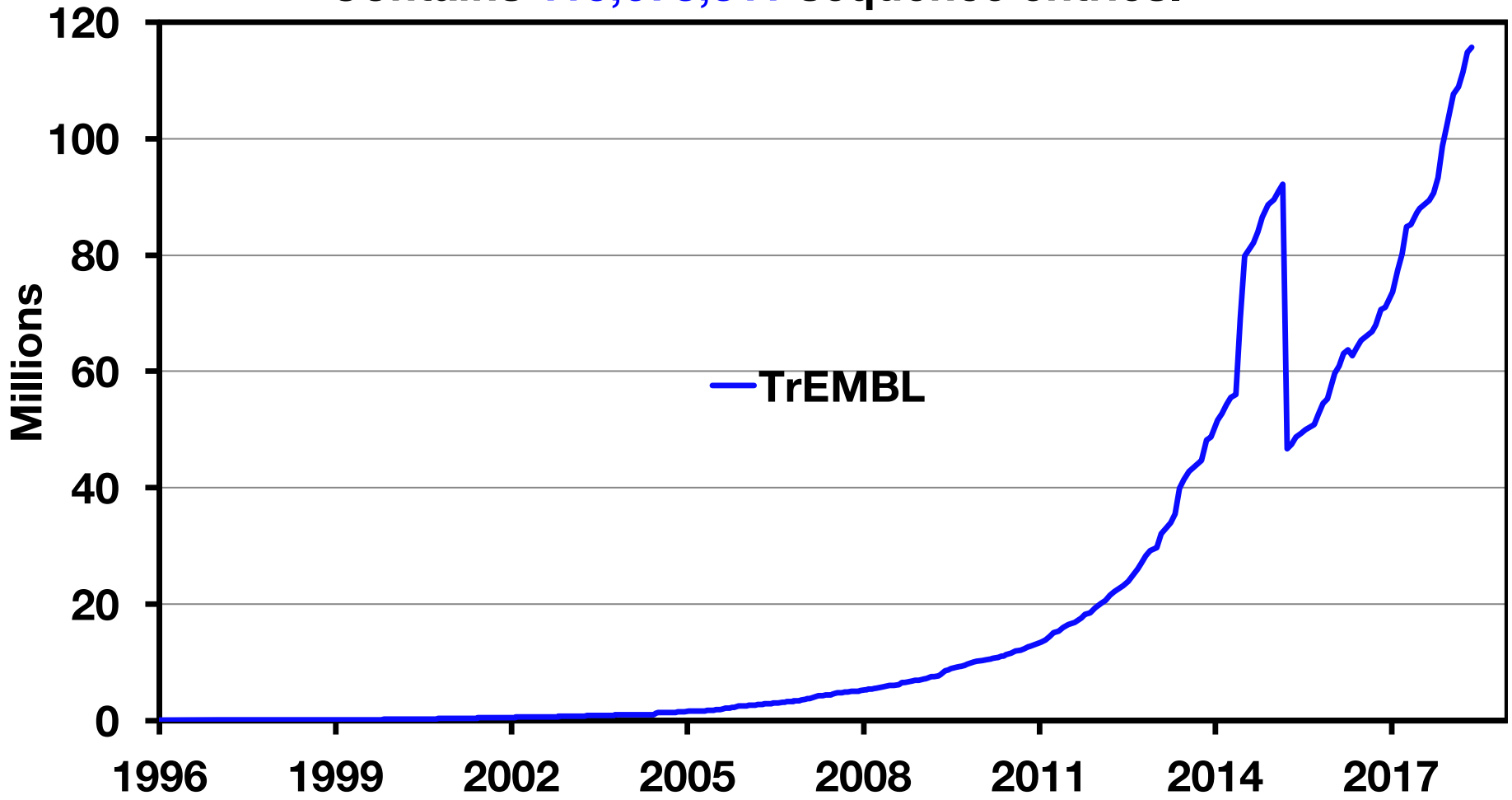
University of Illinois, Urbana-Champaign

Penn State Bioinorganic Workshop
June 2, 2018

The number of protein sequences is “exploding” !

Release 2018_05 of UniProtKB/TrEMBL

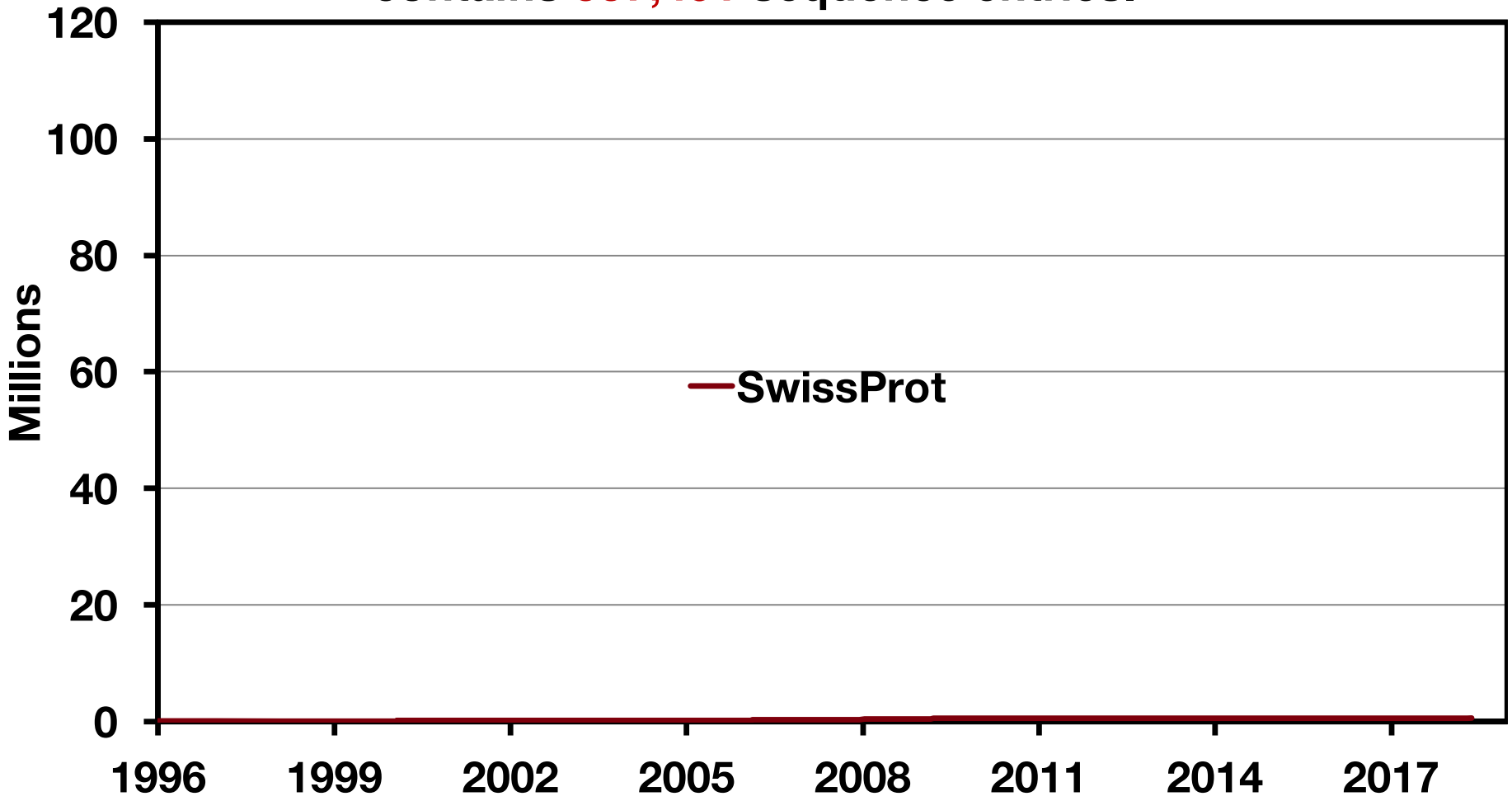
Contains 115,678,811 sequence entries.



But the number of curated annotations is lagging !

Release 2018_05 of UniProtKB/SwissProt

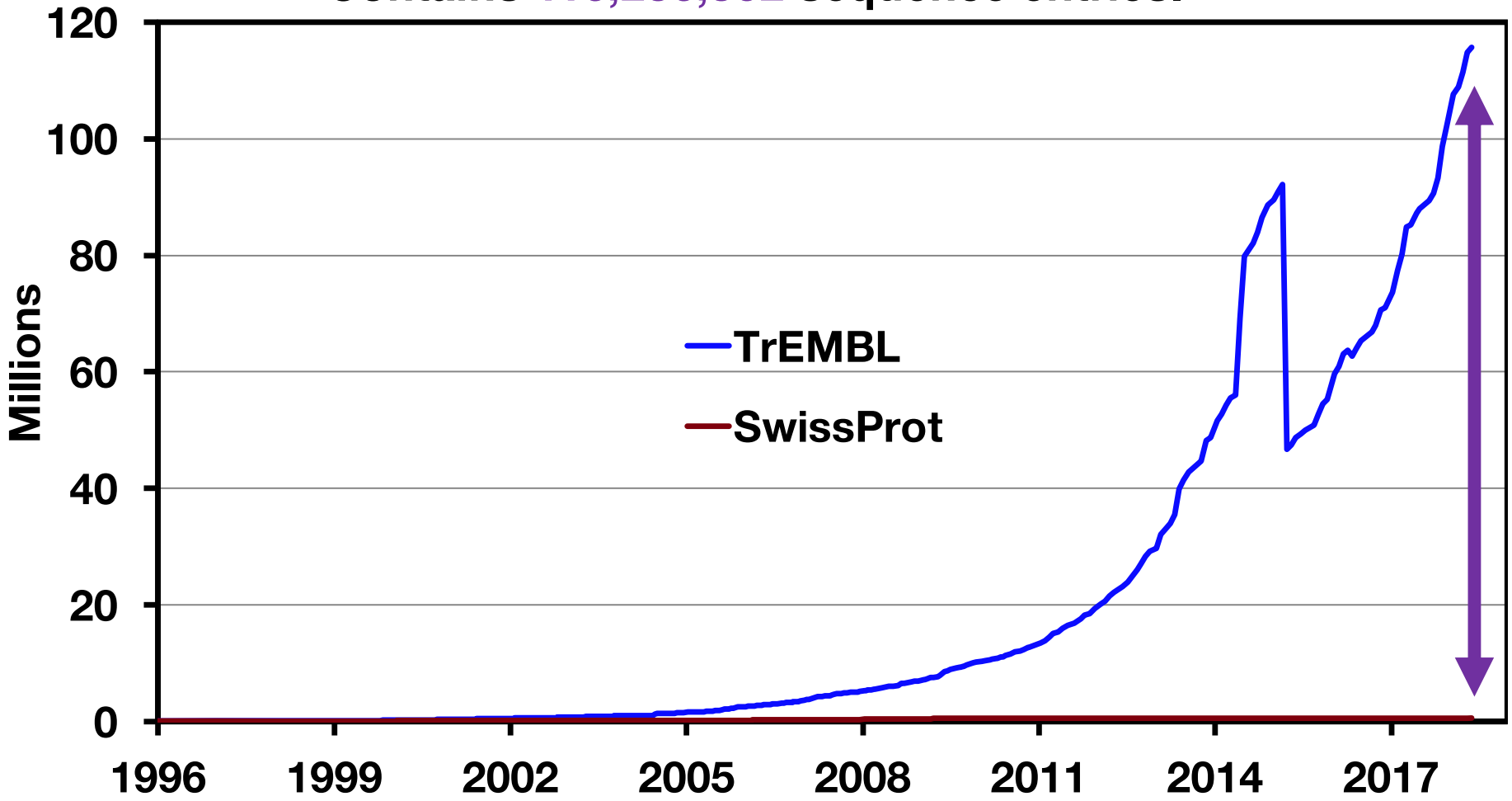
contains 557,491 sequence entries.



Perhaps 50% have unknown or uncertain functions

Release 2018_05 of UniProtKB

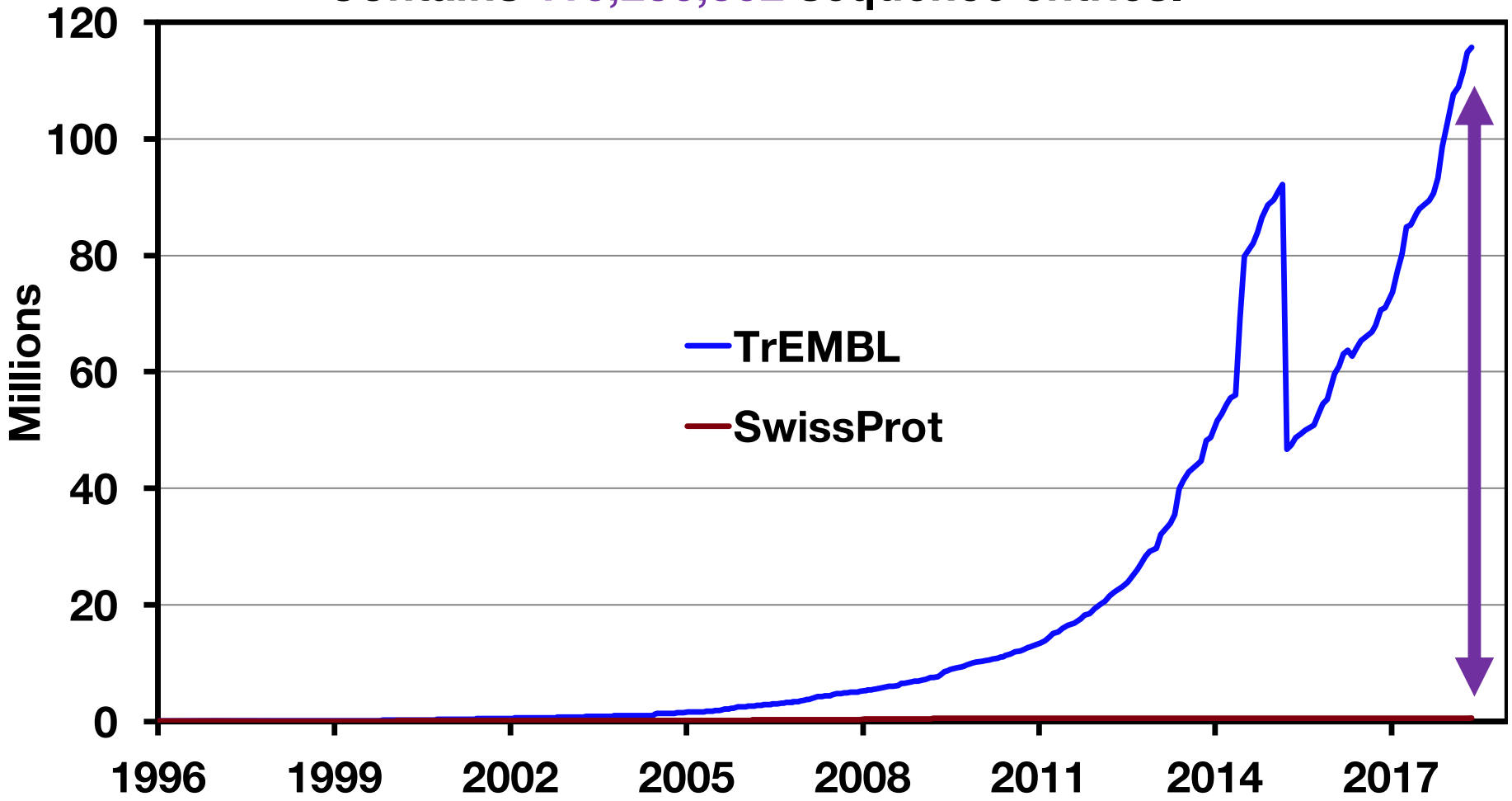
Contains 116,236,302 sequence entries.



How do we solve this problem ?

Release 2018_05 of UniProtKB

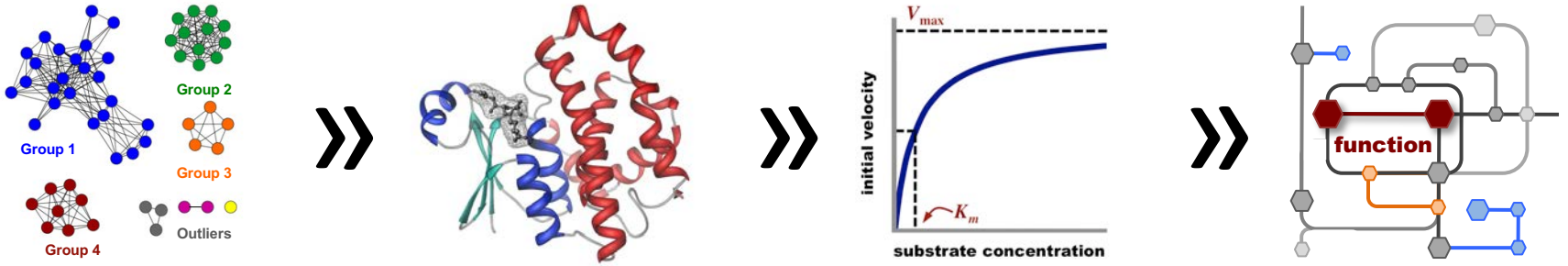
Contains 116,236,302 sequence entries.



Requires large-scale bioinformatics, computational, and experimental tools and strategies to use them

U54GM093342: Enzyme Function Initiative

bioinformatics | *x-ray / computation* | *enzymology* | *biology*



sequence » **structure** » **reaction** » **function**

[Albert Einstein](#)
Steve Almo

[Boston University](#)
Karen Allen

[EMBL-EBI](#)
Alex Bateman

[University of Illinois](#)
John Gerlt
John Cronan

[Pennsylvania State](#)
Squire Booker

[Sanford-Burnham MRI](#)
Andrei Osterman
Dmitry Rodionov

[UCSF](#)
Matthew Jacobson
Andrej Sali
Brian Shoichet

[University of Utah](#)
C. Dale Poulter

[University of Virginia](#)
Wladek Minor

"Genomic Enzymology" Web Tools

Biochemistry

Perspective

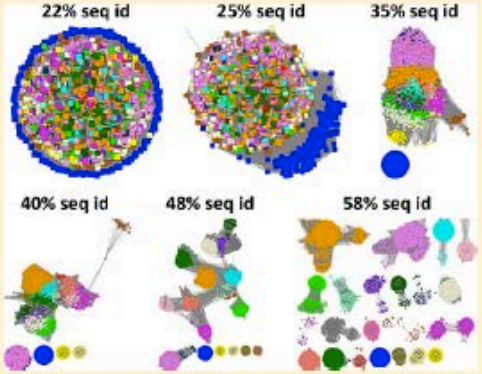
pubs.acs.org/biochemistry

Genomic Enzymology: Web Tools for Leveraging Protein Family Sequence–Function Space and Genome Context to Discover Novel Functions

John A. Gerlt*

Departments of Biochemistry and Chemistry, Institute for Genomic Biology, University of Illinois, Urbana-Champaign Urbana, Illinois 61801, United States

ABSTRACT: The exponentially increasing number of protein and nucleic acid sequences provides opportunities to discover novel enzymes, metabolic pathways, and metabolites/natural products, thereby adding to our knowledge of biochemistry and biology. The challenge has evolved from generating sequence information to mining the databases to integrating and leveraging the available information, i.e., the availability of "genomic enzymology" web tools. Web tools that allow identification of biosynthetic gene clusters are widely used by the natural products/synthetic biology community, thereby facilitating the discovery of novel natural products and the enzymes responsible for their biosynthesis. However, many novel enzymes with interesting mechanisms participate in uncharacterized small-molecule metabolic pathways; their discovery and functional characterization also can be accomplished by leveraging information in protein and nucleic acid databases. This Perspective focuses on two genomic enzymology web tools that assist the discovery novel metabolic pathways: (1) Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST) for generating sequence similarity networks to visualize and analyze sequence–function space in protein families and (2) Enzyme Function Initiative-Genome Neighborhood Tool (EFI-GNT) for generating genome neighborhood networks to visualize and analyze the genome context in microbial and fungal genomes. Both tools have been adapted to other applications to facilitate target selection for enzyme discovery and functional characterization. As the natural products community has demonstrated, the enzymology community needs to embrace the essential role of web tools that allow the protein and genome sequence databases to be leveraged for novel insights into enzymological problems.



***Biochemistry* 2017, 56, 4293-4308**

Genomic enzymology strategy for functional assignment

1. **Explore sequence function-space in a protein family, segregating family into isofunctional clusters using known functions (SwissProt, literature)**

   Sequence similarity networks (SSNs) using EFI-EST

2. **Predict functions of isofunctional clusters in novel metabolic pathways using genome context (types of reactions enabled by proximal enzymes)**

   Genome neighborhood networks (GNNs) using EFI-GNT

3. **Prioritize clusters for functional characterization by (metagenome) abundance**

   Chemically guided functional profiling (CGFP)

Today

1. **Target selection using sequence similarity networks:**
   EFI-EST web tool

2. **Pathway discovery using genome neighborhood networks:**
   EFI-GNT web tool

3. **Example: using SSNs and GNNs to discover catabolic pathways for D-apiose**

4. **Target prioritization using chemically guided functional profiling: CGFP-ShortBRED beta tool**

Today

1. **Target selection using sequence similarity networks:**
   EFI-EST web tool

2. **Pathway discovery using genome neighborhood networks:**
   EFI-GNT web tool

3. **Example: using SSNs and GNNs to discover catabolic pathways for D-apiose**

4. **Target prioritization using chemically guided functional profiling: CGFP-ShortBRED beta tool**

Sequence Similarity Networks (SSNs): Atkinson et al.

OPEN ACCESS Freely available online

PLoS one

Using Sequence Similarity Networks for Visualization of Relationships Across Diverse Protein Superfamilies

Holly J. Atkinson[1,2], John H. Morris[3], Thomas E. Ferrin[2,3,4], Patricia C. Babbitt[2,3,4]*

1 Graduate Program in Biological and Medical Informatics, University of California San Francisco, San Francisco, California, United States of America, **2** Institute for Quantitative Biosciences, University of California San Francisco, San Francisco, California, United States of America, **3** Department of Pharmaceutical Chemistry, University of California San Francisco, San Francisco, California, United States of America, **4** Department of Biopharmaceutical Sciences, University of California San Francisco, San Francisco, California, United States of America

Abstract

The dramatic increase in heterogeneous types of biological data—in particular, the abundance of new protein sequences—requires fast and user-friendly methods for organizing this information in a way that enables functional inference. The most widely used strategy to link sequence or structure to function, homology-based function prediction, relies on the fundamental assumption that sequence or structural similarity implies functional similarity. New tools that extend this approach are still urgently needed to associate sequence data with biological information in ways that accommodate the real complexity of the problem, while being accessible to experimental as well as computational biologists. To address this, we have examined the application of sequence similarity networks for visualizing functional trends across protein superfamilies from the context of sequence similarity. Using three large groups of homologous proteins of varying types of structural and functional diversity—GPCRs and kinases from humans, and the crotonase superfamily of enzymes—we show that overlaying networks with orthogonal information is a powerful approach for observing functional themes and revealing outliers. In comparison to other primary methods, networks provide both a good representation of group-wise sequence similarity relationships and a strong visual and quantitative correlation with phylogenetic trees, while enabling analysis and visualization of much larger sets of sequences than trees or multiple sequence alignments can easily accommodate. We also define important limitations and caveats in the application of these networks. As a broadly accessible and effective tool for the exploration of protein superfamilies, sequence similarity networks show great potential for generating testable hypotheses about protein structure-function relationships.

Citation: Atkinson HJ, Morris JH, Ferrin TE, Babbitt PC (2009) Using Sequence Similarity Networks for Visualization of Relationships Across Diverse Protein Superfamilies. PLoS ONE 4(2): e4345. doi:10.1371/journal.pone.0004345

Editor: I. King Jordan, Georgia Institute of Technology, United States of America

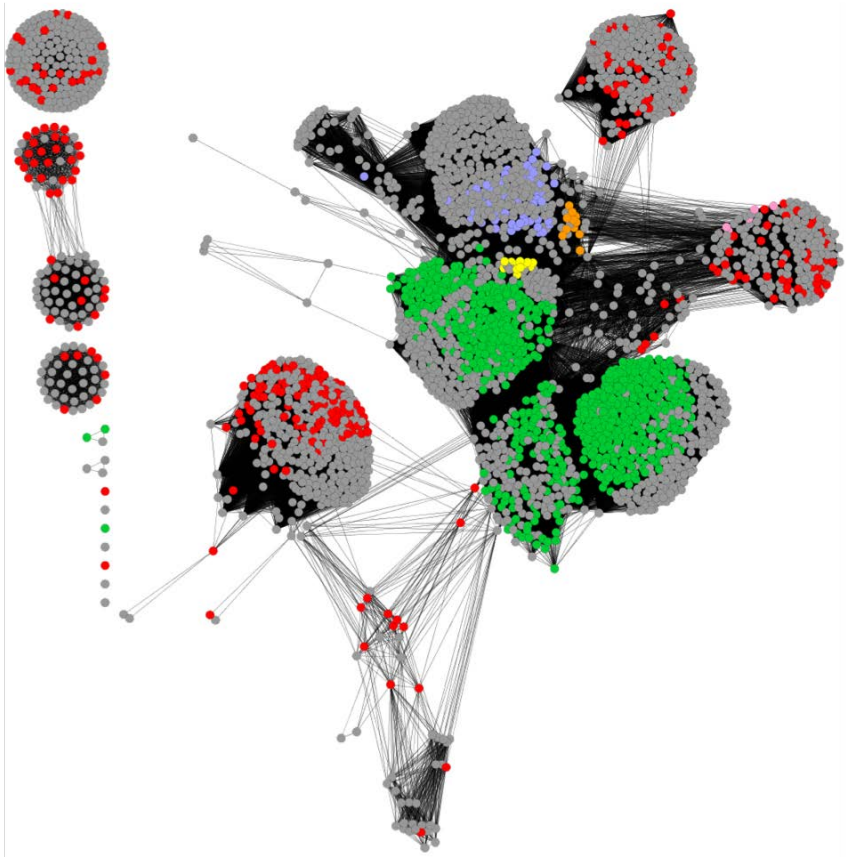
Received September 10, 2008; **Accepted** December 10, 2008; **Published** February 3, 2009

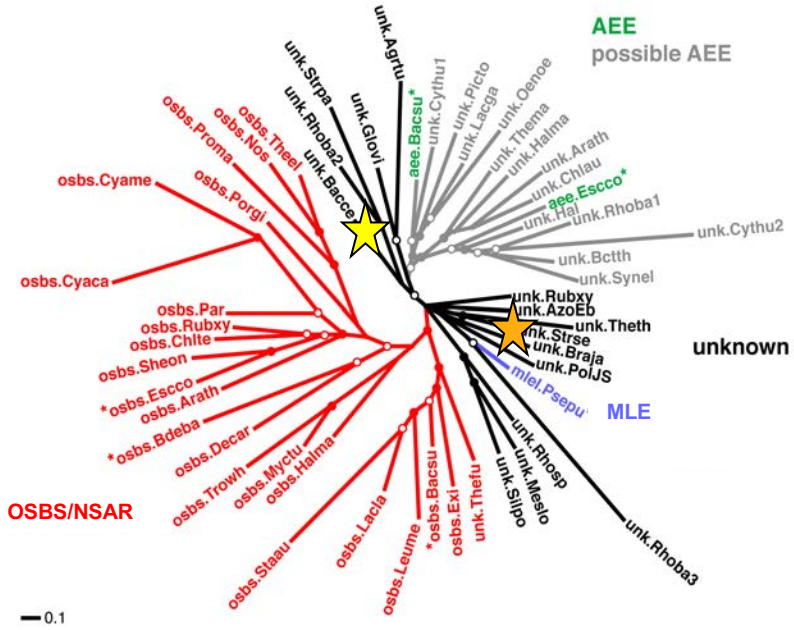
Funding: This work was supported by NIH grant GM60595 and NSF grant DBI 0640476 to P.C.B. and P41 RR01081 to T.E.F.. H.J.A. received support from NIH grant T32 GM067547. Initial exploration of sequence similarity networks used the enolases and amidohydrolases superfamilies as example data sets, and was supported by P01 GM071790 to P.C.B.. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: babbitt@cgl.ucsf.edu

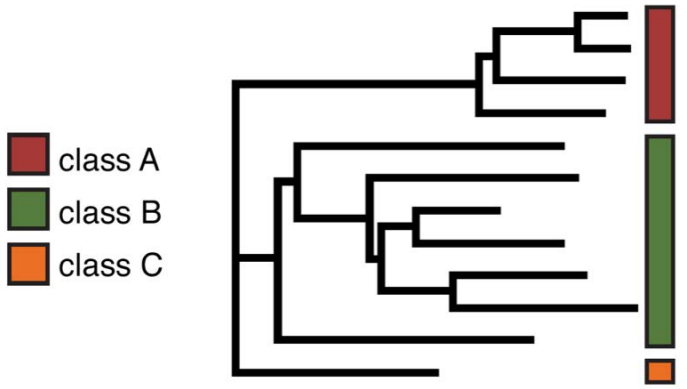
Sequence similarity networks (SSNs) vs. dendrograms: large-scale visualization tool for target selection





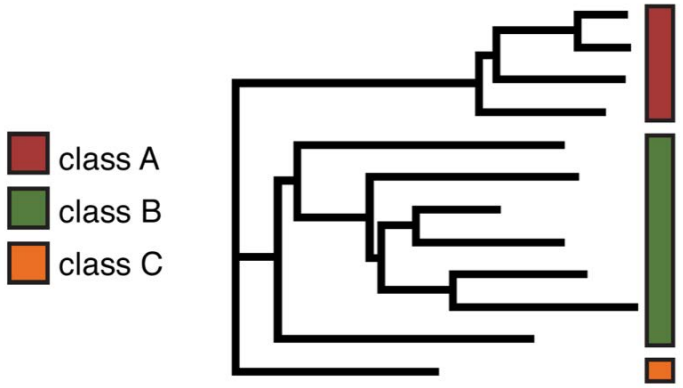
Families are easier to visualize and connect with SSNs, so function predictions are easier to formulate and explore

Dendrograms/trees for sequence relationships



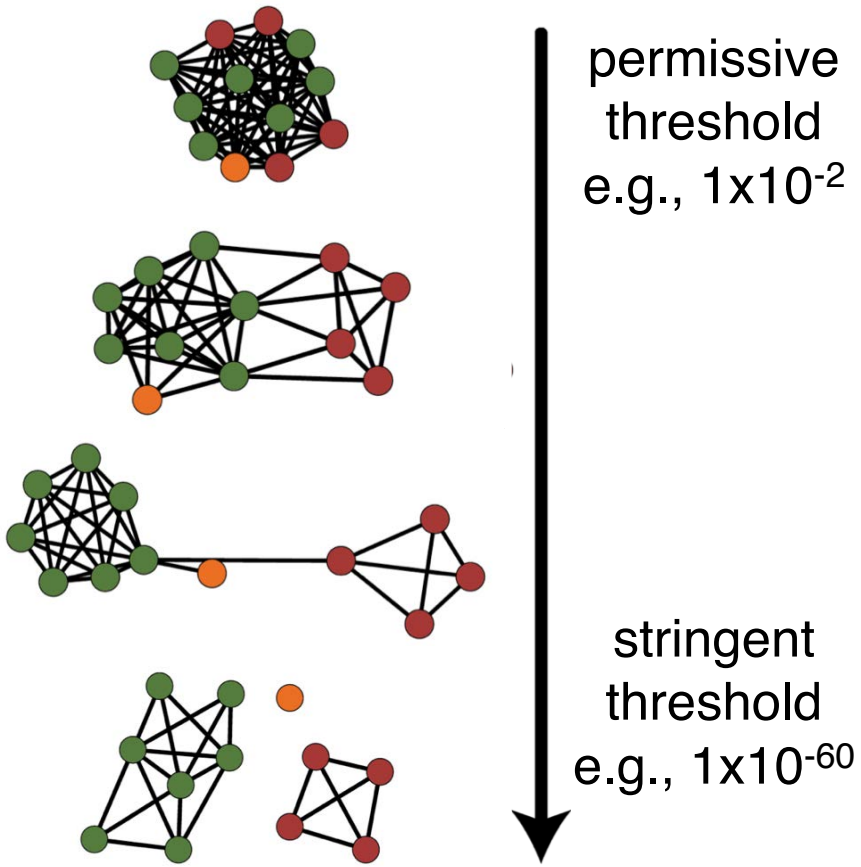
Babbitt et al., PLoS One **2009**, *4*, e4345

Connectivity: structure-based multiple sequence alignments (MSAs)



Babbitt et al., PLoS One **2009**, *4*, e4345

Sequence similarity networks

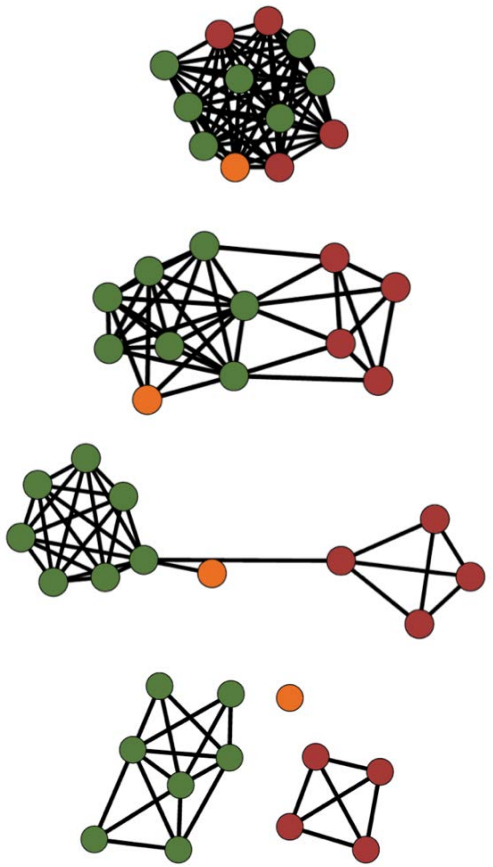


node (circle) = sequence

edge (line) = connection less than a user-defined similarity score (e-value)

Babbitt et al., PLoS One **2009**, *4*, e4345

Connectivity: all-by-all BLASTP e-values



permissive threshold
e.g., $1x10^{-2}$

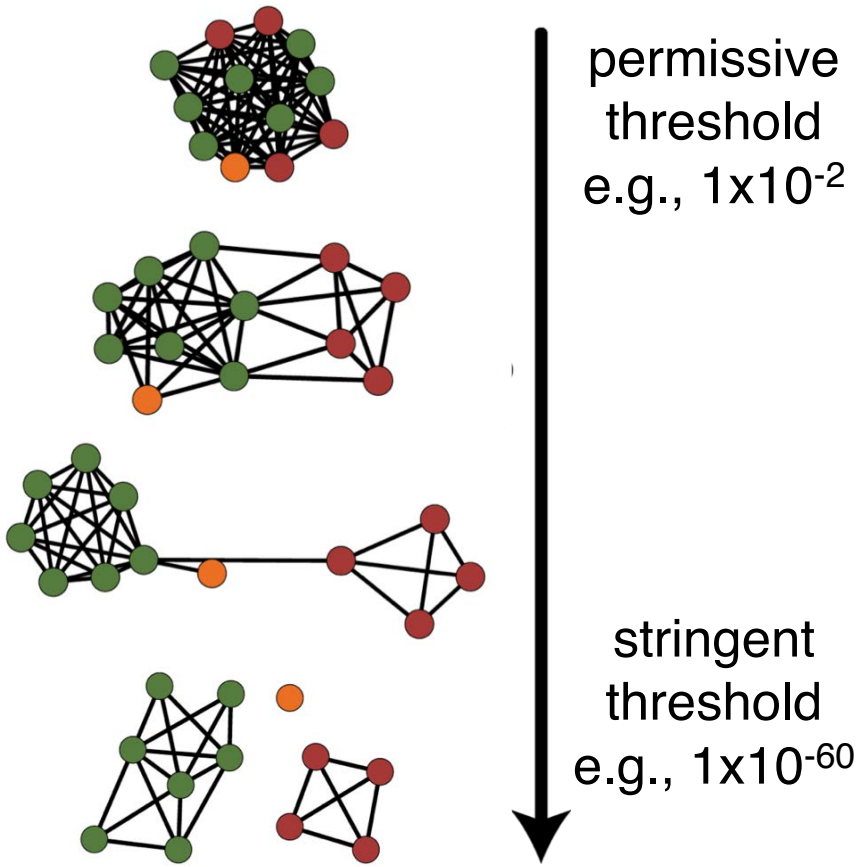
stringent threshold
e.g., $1x10^{-60}$

node (circle) = sequence

edge (line) = connection less than a user-defined similarity score (e-value)

Babbitt et al., PLoS One **2009**, *4*, e4345

Faster to calculate than MSAs

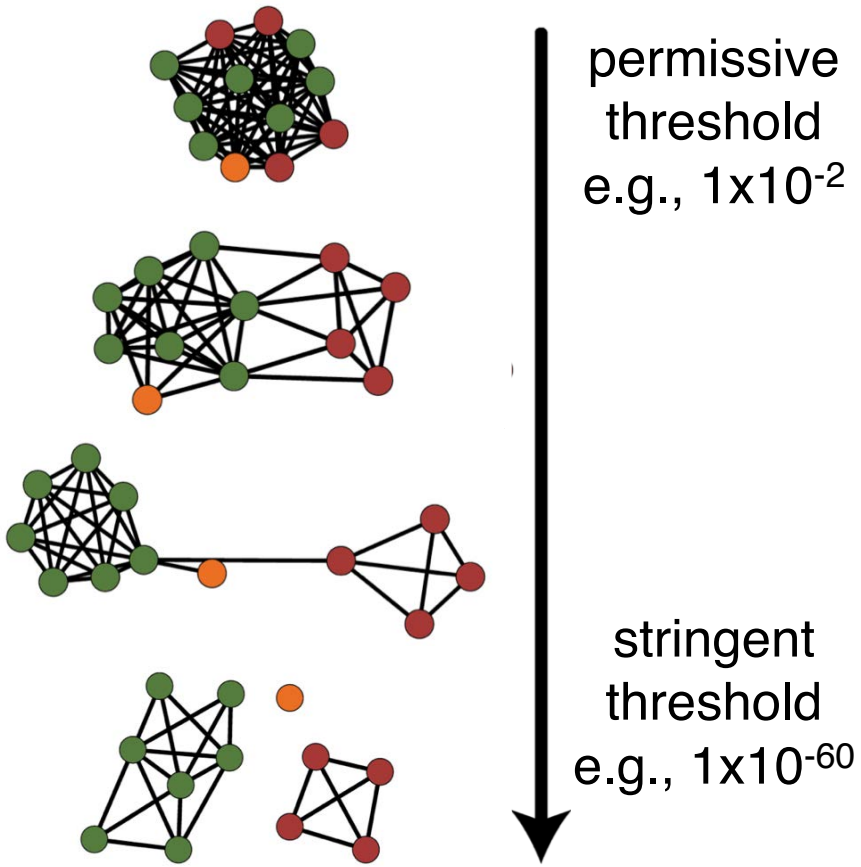


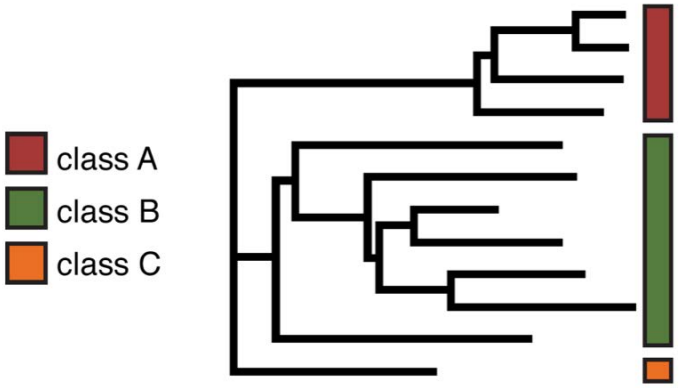
node (circle) = sequence

edge (line) = connection less than a user-defined similarity score (e-value)

Babbitt et al., PLoS One **2009**, *4*, e4345

Qualitatively similar results



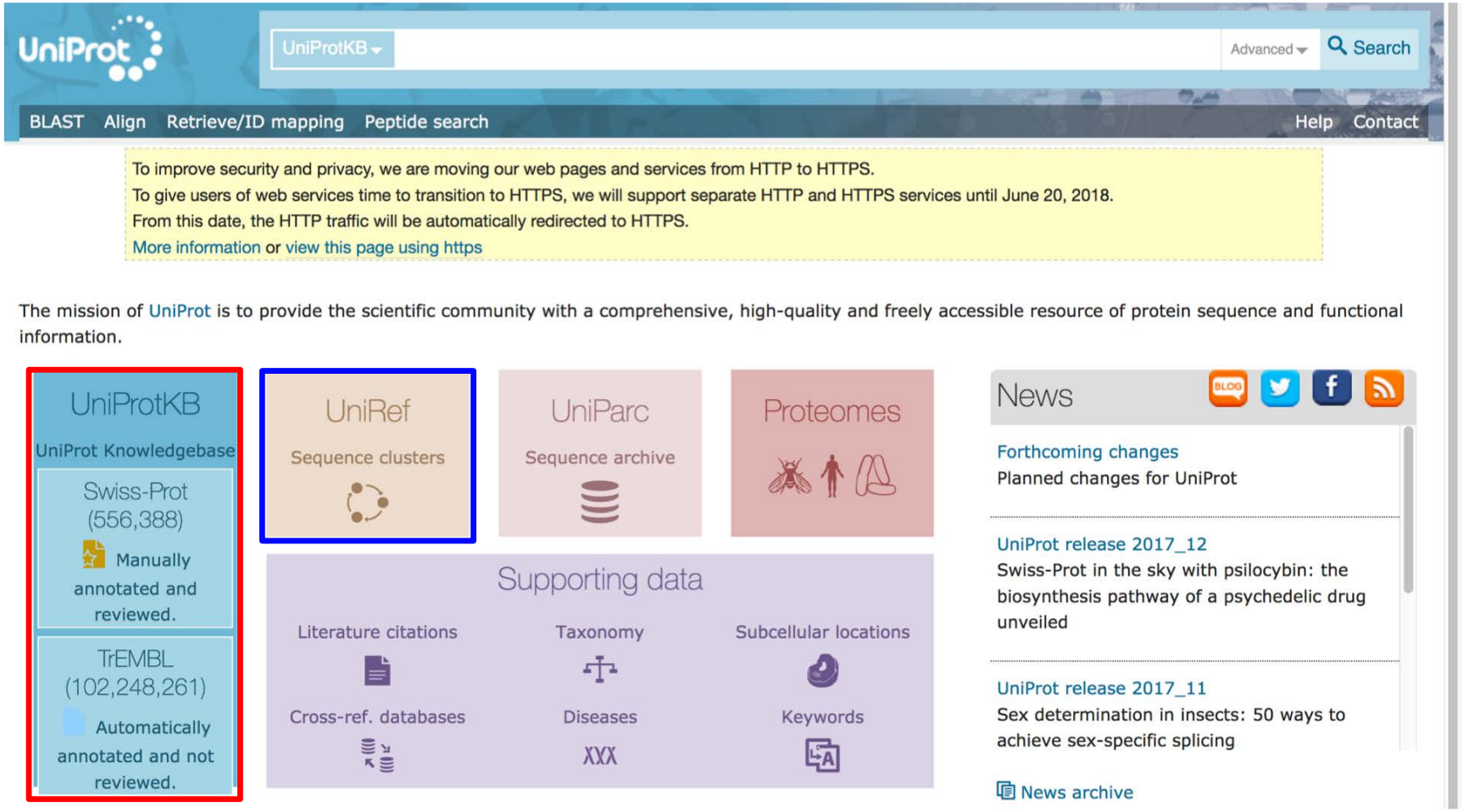


node (circle) = sequence

edge (line) = connection less than a user-defined similarity score (e-value)

Babbitt et al., PLoS One **2009**, *4*, e4345

EFI-EST uses sequences from UniProt (EMBL/EBI)



http://www.uniprot.org/

UniProt databases used by EFI-EST

UniProt Knowledgebase (UniProtKB): curated protein information, including function, classification, and cross-reference.

UniProtKB/Swiss-Prot: manually annotated

UniProtKB/TrEMBL: automatically annotated

UniProt Reference Clusters (UniRef): clustered sets of sequences from UniProtKB and UniProt Archive records to obtain complete coverage of sequence space at several resolutions while hiding redundant sequence.

UniRef90: 90% sequence identity over 80% of the length of the longest sequence (used by Option B of EFI-EST). On average, a two-fold reduction in number of sequences, providing a four-fold reduction in running time.

Why UniProt, not NCBI? Annotations can be changed

Contact

Submit updates or corrections to UniProt

› We are **actively seeking** any type of updates and/or corrections whether they have been published or not.

› Please see the guidelines for submitting updates.

Other ways to contact us...

Send us general questions and suggestions
Submit new protein sequence data
Email us on help@uniprot.org

Name :	
E-mail :	
Subject :	UniProtKB/TrEMBL D2JA13 entry update r
This is currently an unreviewed entry. We are going to update it and integrate it as a reviewed entry in UniProtKB/Swiss-Prot. Please post your update in the text field below.	
Message :	

https://www.uniprot.org/update

EFI-EST uses families from Pfam and InterPro

Pfam: conserved protein families based a seed alignment of representative sequences that is used to generate a profile hidden Markov model (HMM). Includes families and clans (superfamilies of multiple families).

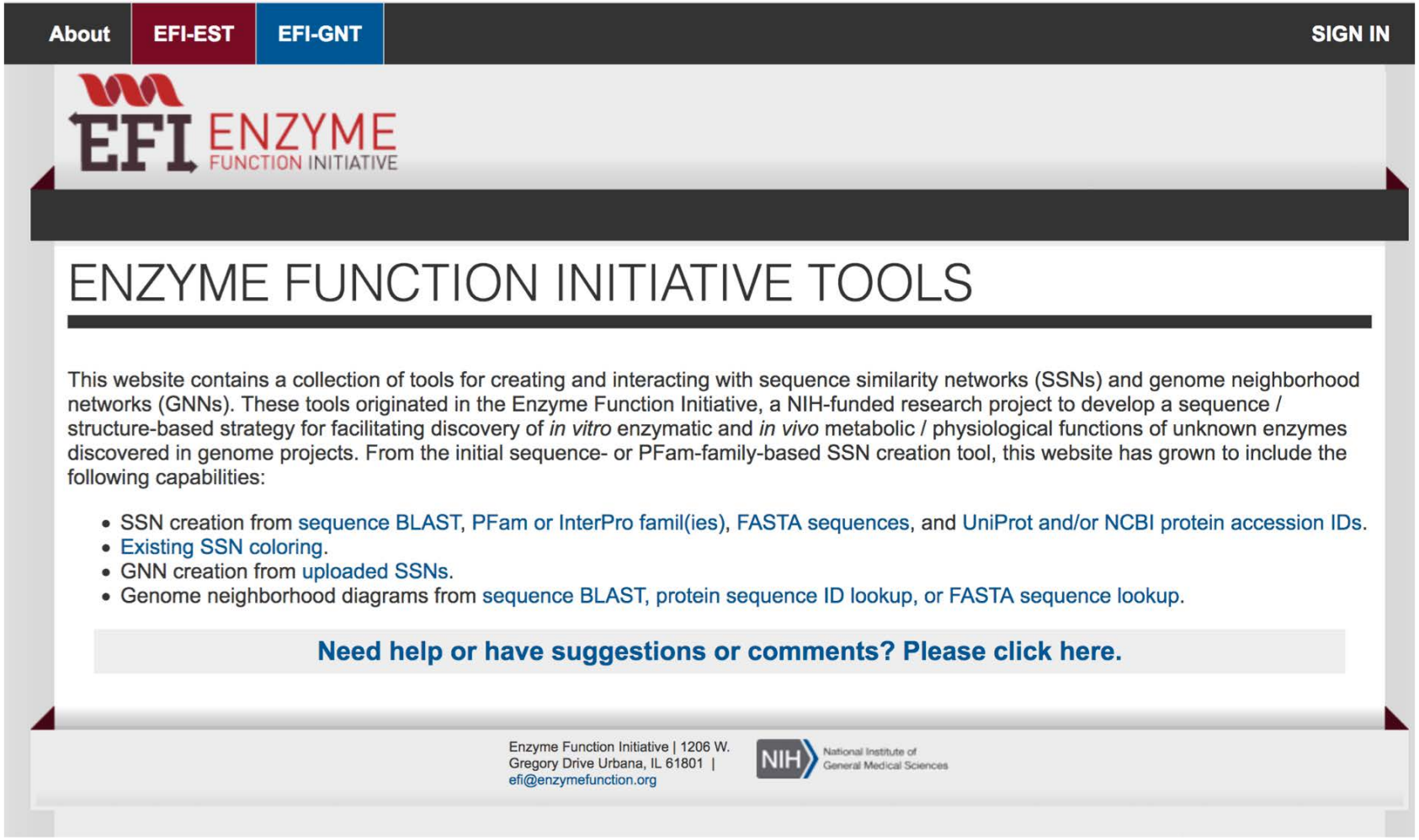
16,712 families in Pfam 31.0 (March 2017)

InterPro: aggregates of families, domains, and sites curated by 14 databases, e.g., Pfam, CATH-Gene3D, ProSite, SFLD, TIGRFAMs, etc.

33,947 domains, families, and sites in InterPro 68 (April 28, 2018)

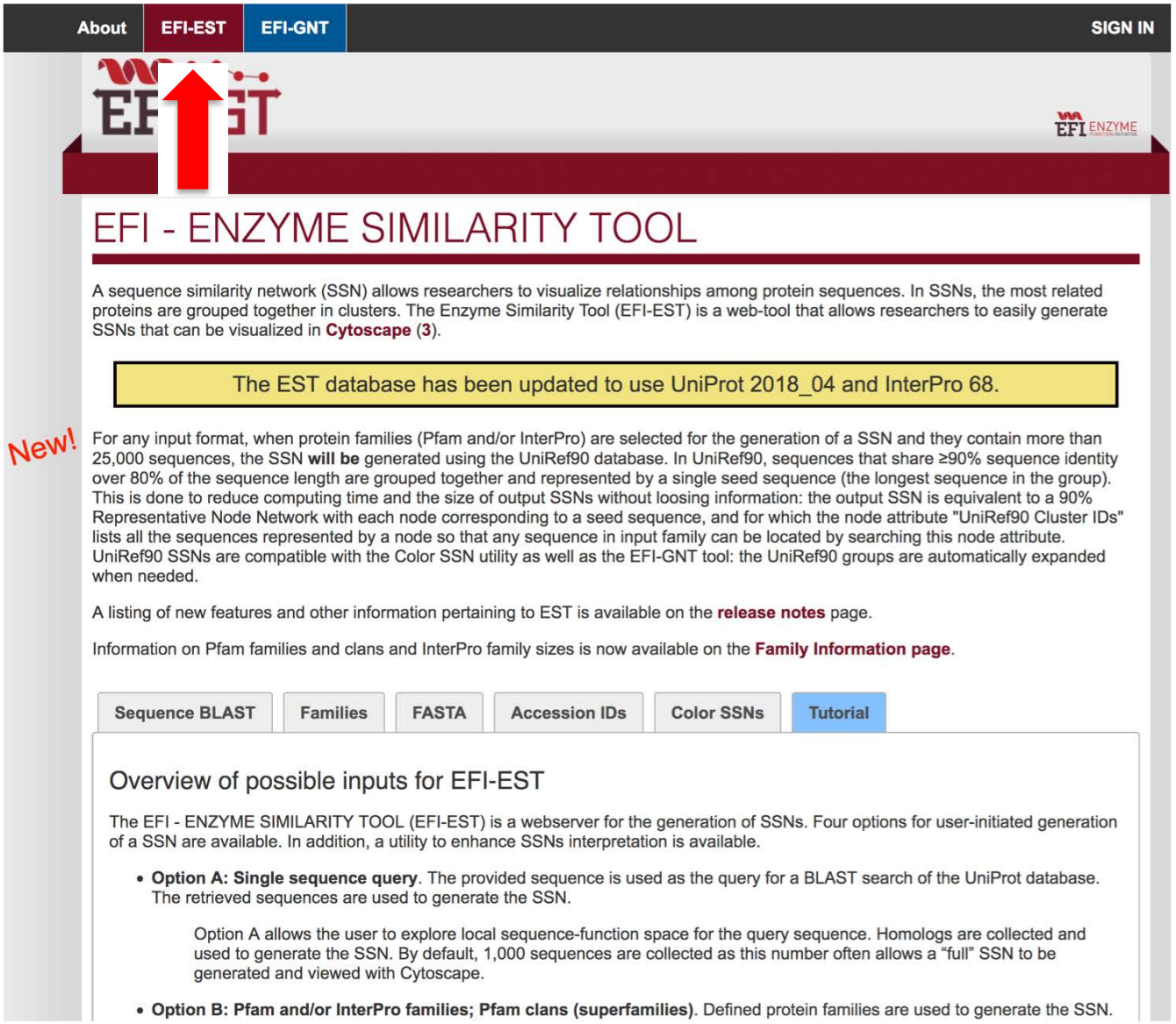
http://pfam.xfam.org/; www.ebi.ac.uk/interpro/

EFI-EST and EFI-GNT Web Tools



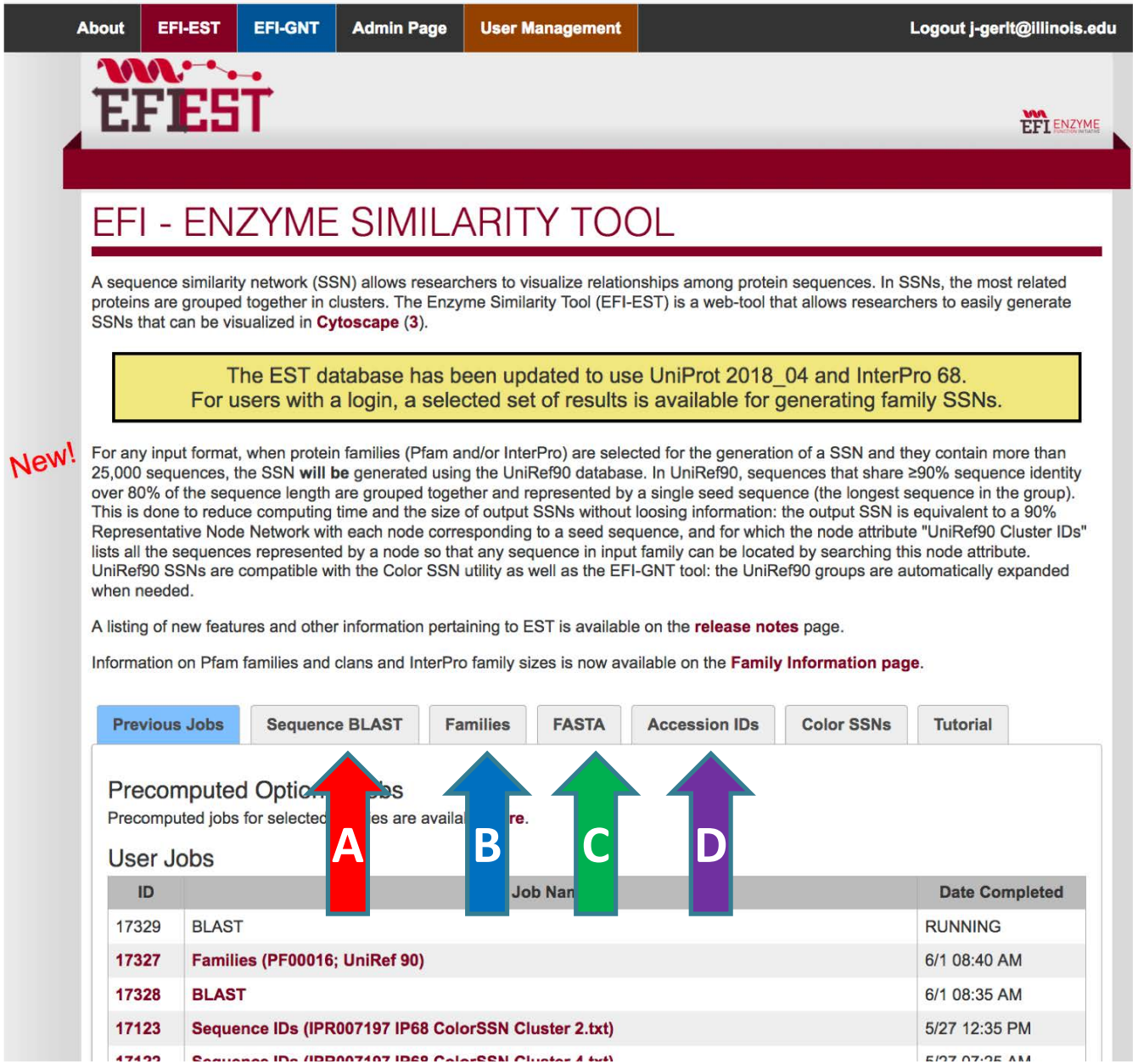
https://efi.igb.illinois.edu/

EFI-EST



https://efi.igb.illinois.edu/efi-est/

Four options for generating SSNs



Four options for generating SSNs

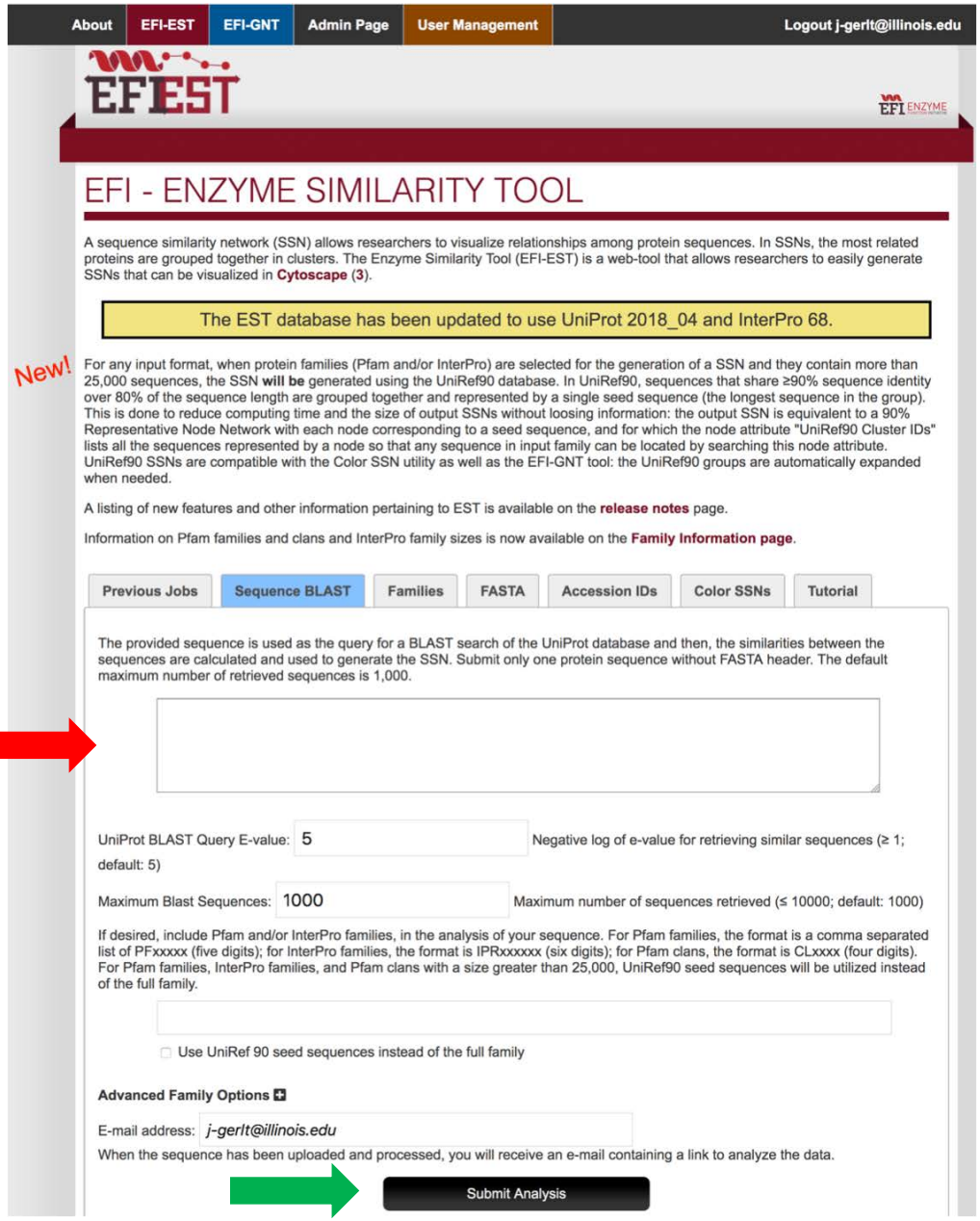
- **Option A:** Homologues in UniProt to user-supplied sequences (collected by BLAST)

- **Option B:** User-supplied Pfam and/or InterPro families from UniProt, using full length sequences or domains

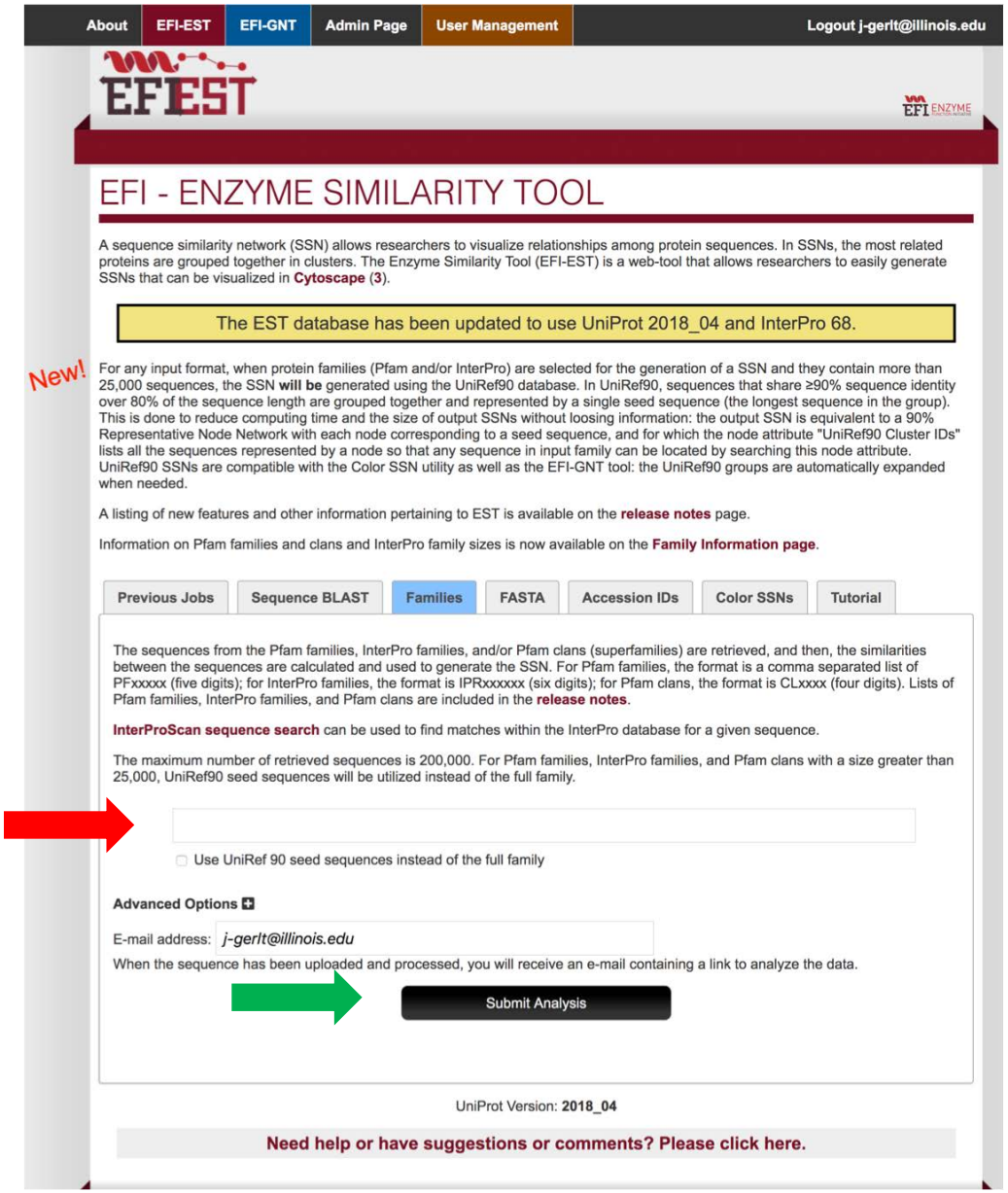
- **Option C:** User-supplied FASTA file; FASTA header can be "read" for **UniProt** and/or **NCBI** IDs

- **Option D:** User-supplied file of **UniProt** and/or **NCBI** IDs

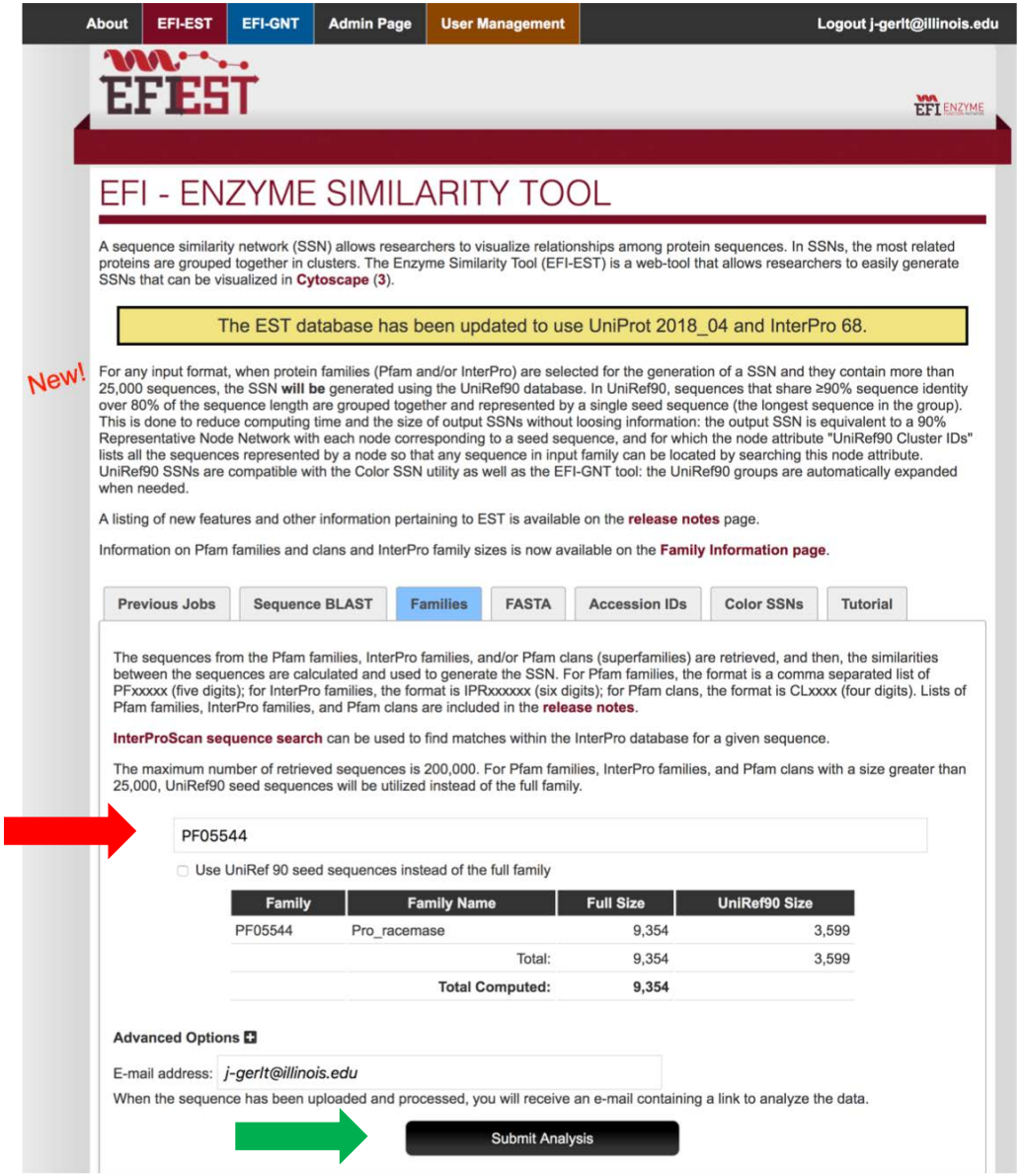
Option A: Homologues collected by BLAST



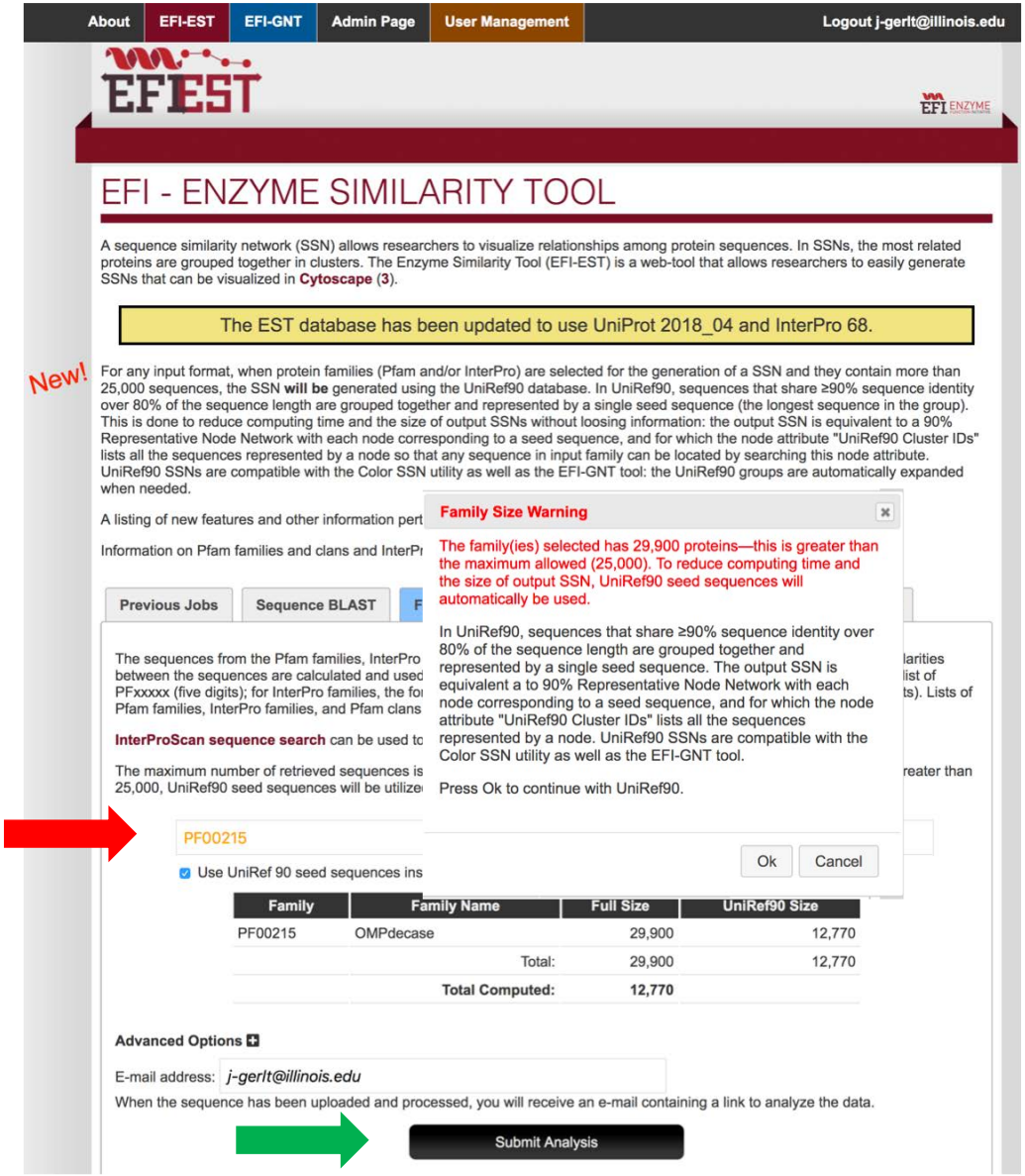
Option B: Pfam and InterPro families, Pfam clans



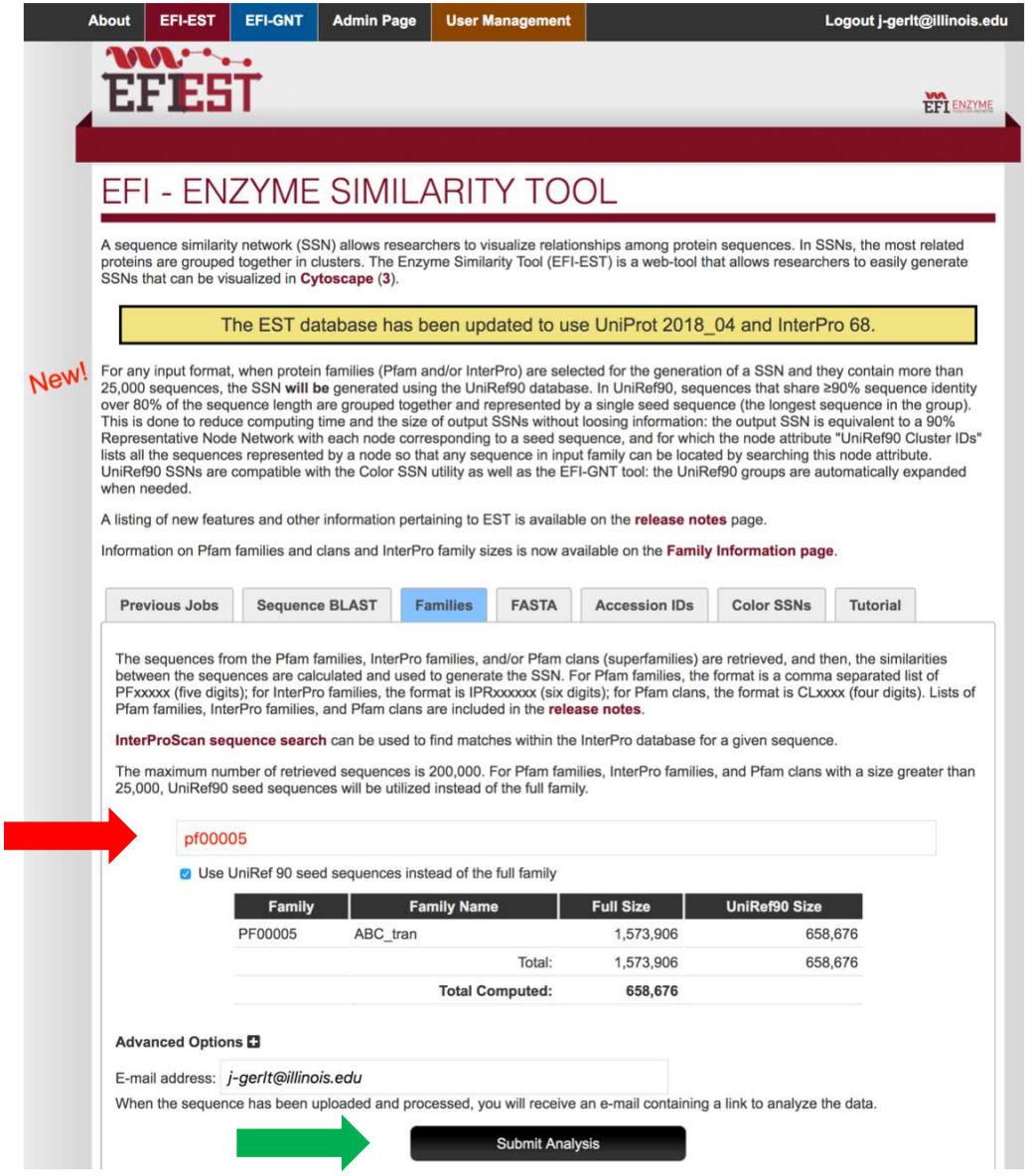
Option B: UniRef for <25K sequences (15,547 Pfams)



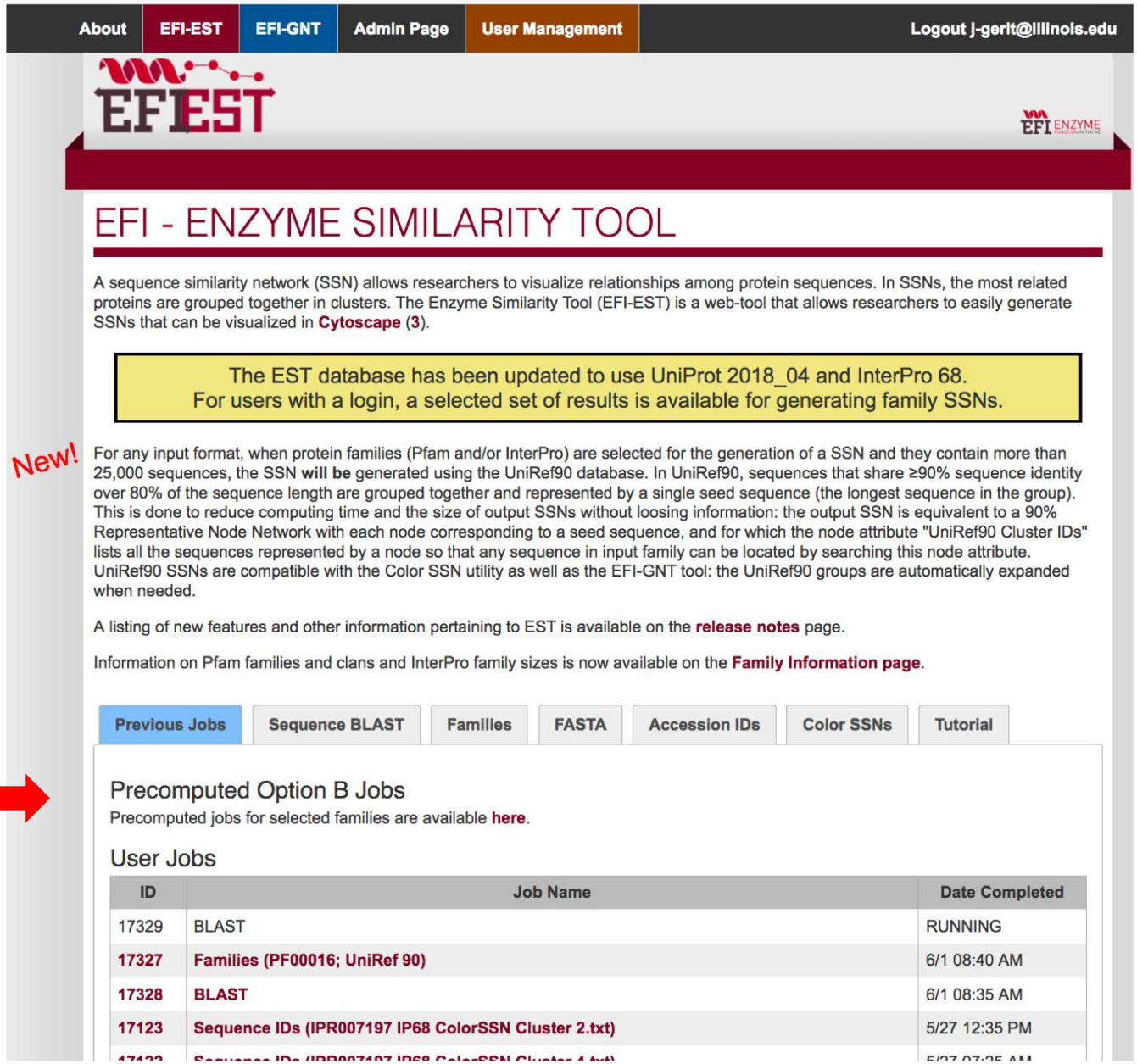
Option B: UniRef90 for ≥ 25K sequences (1,157 Pfams)



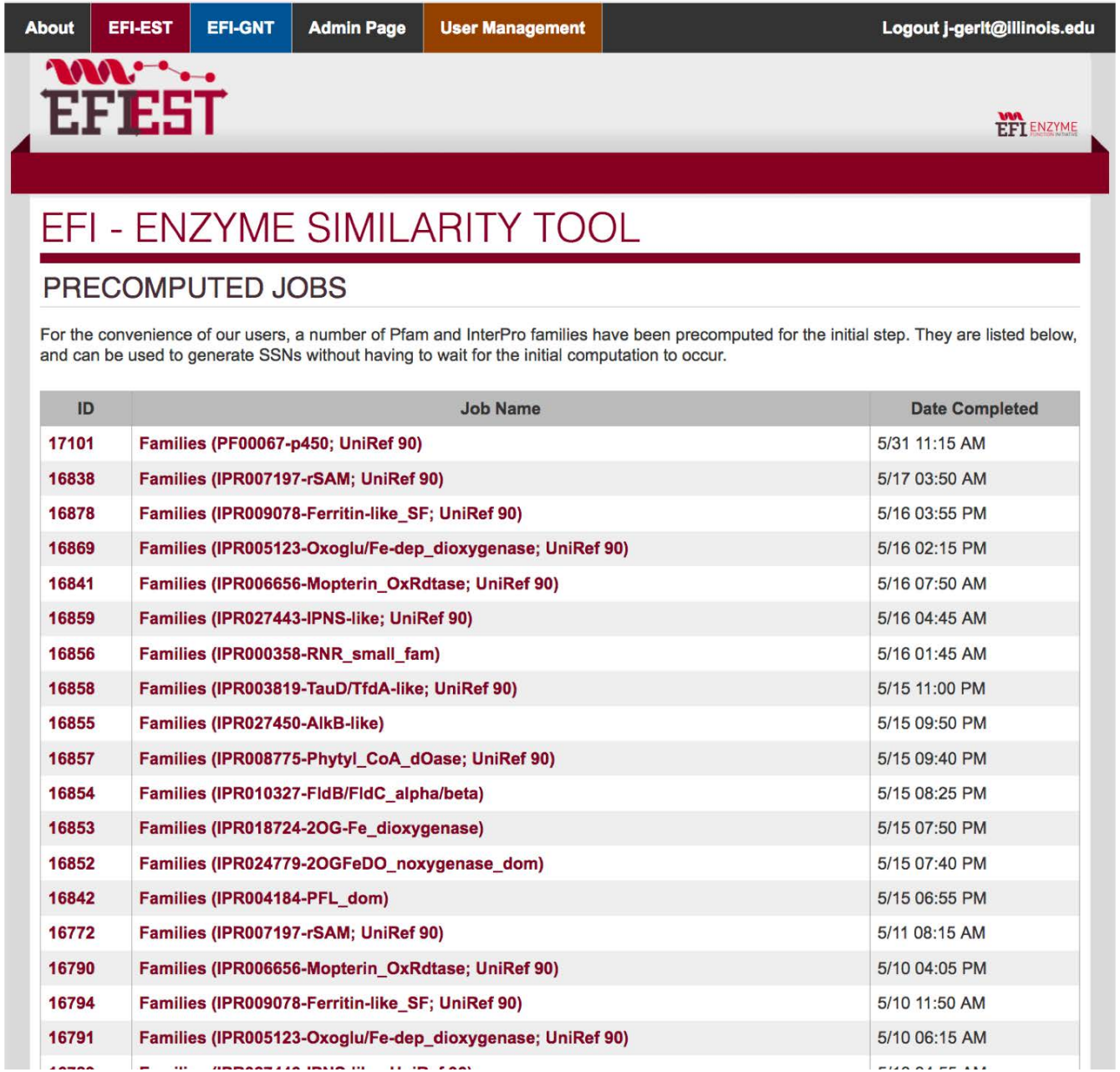
Option B: >200K UniRef90 sequences (8 Pfams)



Precomputed “bioinorganic” Option B jobs



Precomputed “bioinorganic” Option B jobs



About | EFI-EST | EFI-GNT | Admin Page | User Management | Logout j-gerlt@illinois.edu

EFI EST

EFI ENZYME

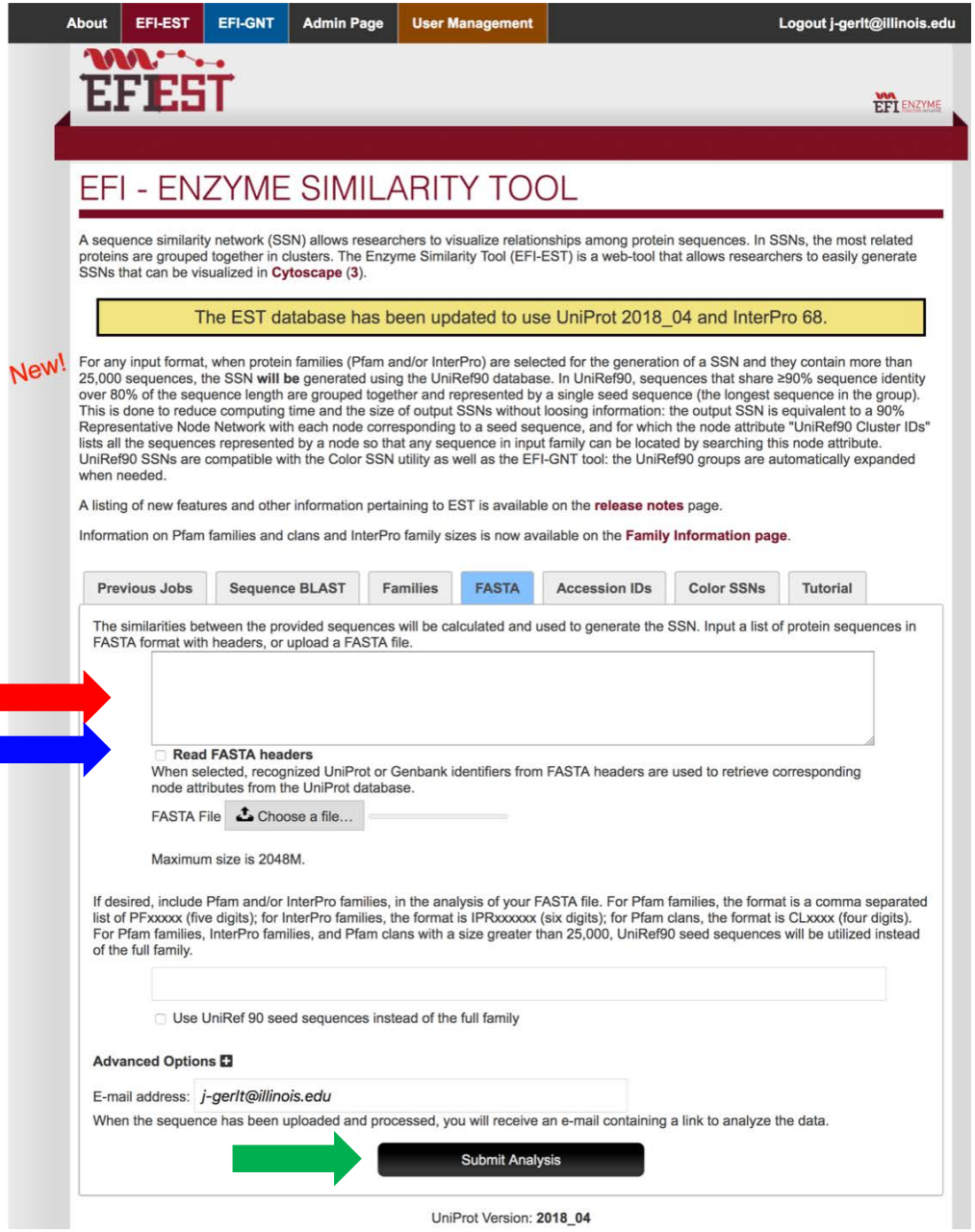
EFI - ENZYME SIMILARITY TOOL

PRECOMPUTED JOBS

For the convenience of our users, a number of Pfam and InterPro families have been precomputed for the initial step. They are listed below, and can be used to generate SSNs without having to wait for the initial computation to occur.

ID	Job Name	Date Completed
17101	Families (PF00067-p450; UniRef 90)	5/31 11:15 AM
16838	Families (IPR007197-rSAM; UniRef 90)	5/17 03:50 AM
16878	Families (IPR009078-Ferritin-like_SF; UniRef 90)	5/16 03:55 PM
16869	Families (IPR005123-Oxoglu/Fe-dep_dioxygenase; UniRef 90)	5/16 02:15 PM
16841	Families (IPR006656-Mopterin_OxRdtase; UniRef 90)	5/16 07:50 AM
16859	Families (IPR027443-IPNS-like; UniRef 90)	5/16 04:45 AM
16856	Families (IPR000358-RNR_small_fam)	5/16 01:45 AM
16858	Families (IPR003819-TauD/TfdA-like; UniRef 90)	5/15 11:00 PM
16855	Families (IPR027450-AlkB-like)	5/15 09:50 PM
16857	Families (IPR008775-Phytyl_CoA_dOase; UniRef 90)	5/15 09:40 PM
16854	Families (IPR010327-FldB/FldC_alpha/beta)	5/15 08:25 PM
16853	Families (IPR018724-2OG-Fe_dioxygenase)	5/15 07:50 PM
16852	Families (IPR024779-2OGFeDO_noxygenase_dom)	5/15 07:40 PM
16842	Families (IPR004184-PFL_dom)	5/15 06:55 PM
16772	Families (IPR007197-rSAM; UniRef 90)	5/11 08:15 AM
16790	Families (IPR006656-Mopterin_OxRdtase; UniRef 90)	5/10 04:05 PM
16794	Families (IPR009078-Ferritin-like_SF; UniRef 90)	5/10 11:50 AM
16791	Families (IPR005123-Oxoglu/Fe-dep_dioxygenase; UniRef 90)	5/10 06:15 AM

Option C: FASTA file; header can be "read" for IDs



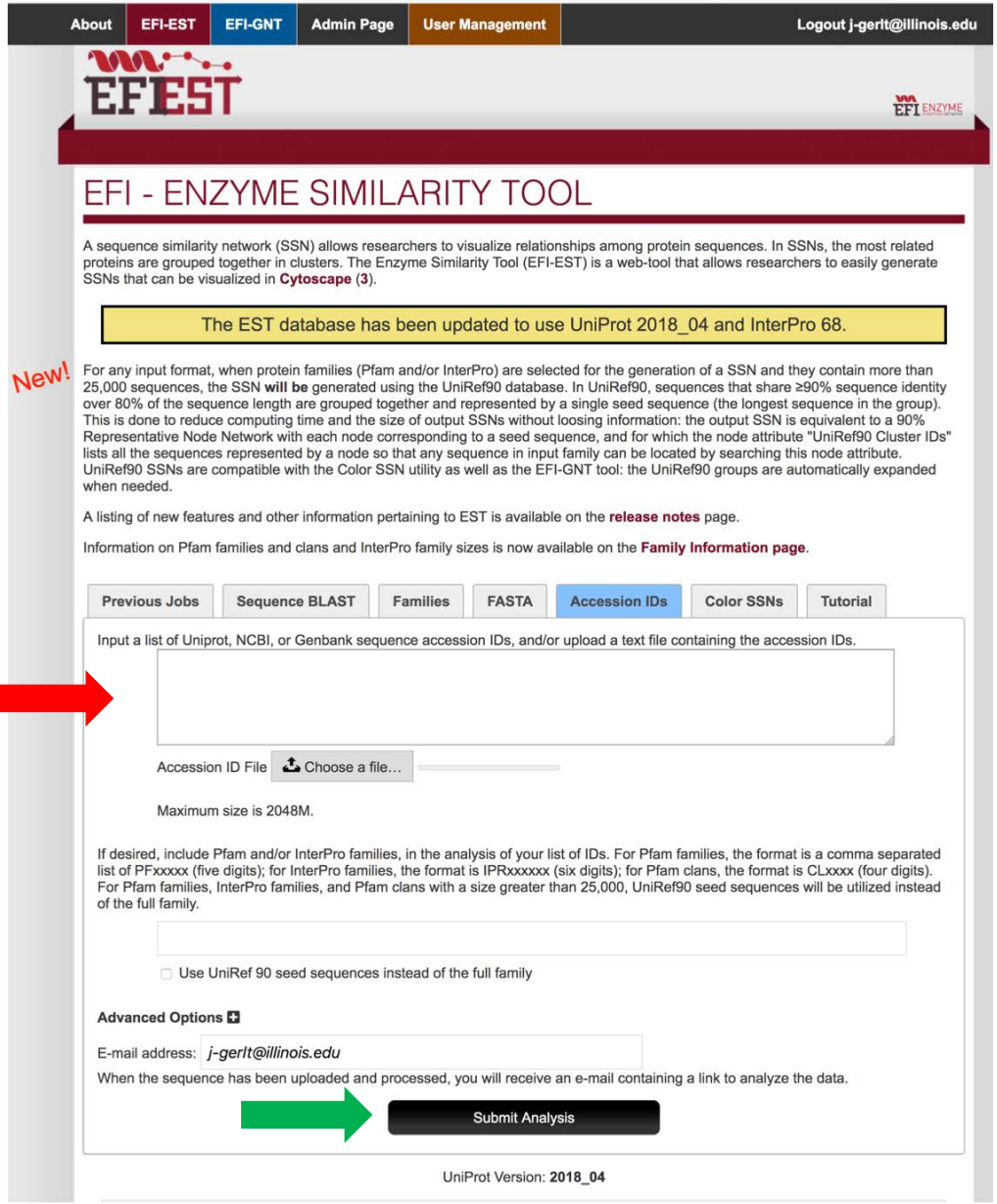
Option C: FASTA header with UniProt IDs

>sp|**A6WXX7**|4HPE2_OCHA4 4-hydroxyproline 2-epimerase 2 OS=Ochrobactrum anthropi (strain ATCC 49188 / DSM 6882 / JCM 21032 / NBRC 15819 / NCTC 12168) GN=Oant_1111 PE=1 SV=1
MARHSFFCVDGHTCGNPVRLVAGGGPNLEGSTMMEKRAHFLREYDWIRTGLMFEPRGHDM
MSGSILYPPTRPDCDVAVLFIETSGCLPMCGHGTIGTVTMAIEQGLVTPKTPGKLNLDTP
AGLVAIEYEQNGQYVERVRLTNVPAFLYAEGLEVECPDLGNLKVDVAYGGNFYAIVEPQE
NYTDMEDYSALQLIAWSPILRERLNEKYKFQHPLLPDINRLSHILWTGKPKHPEAHARNA
VFYGDKAIDRSPCGTGTSARMAQLAAKGKLKPGDEFVHESIIGSLFHGRVERATEVVGQD
RTLPAIIPSIAGWARMTGYNTIFIDDRDPFAHGFTVA

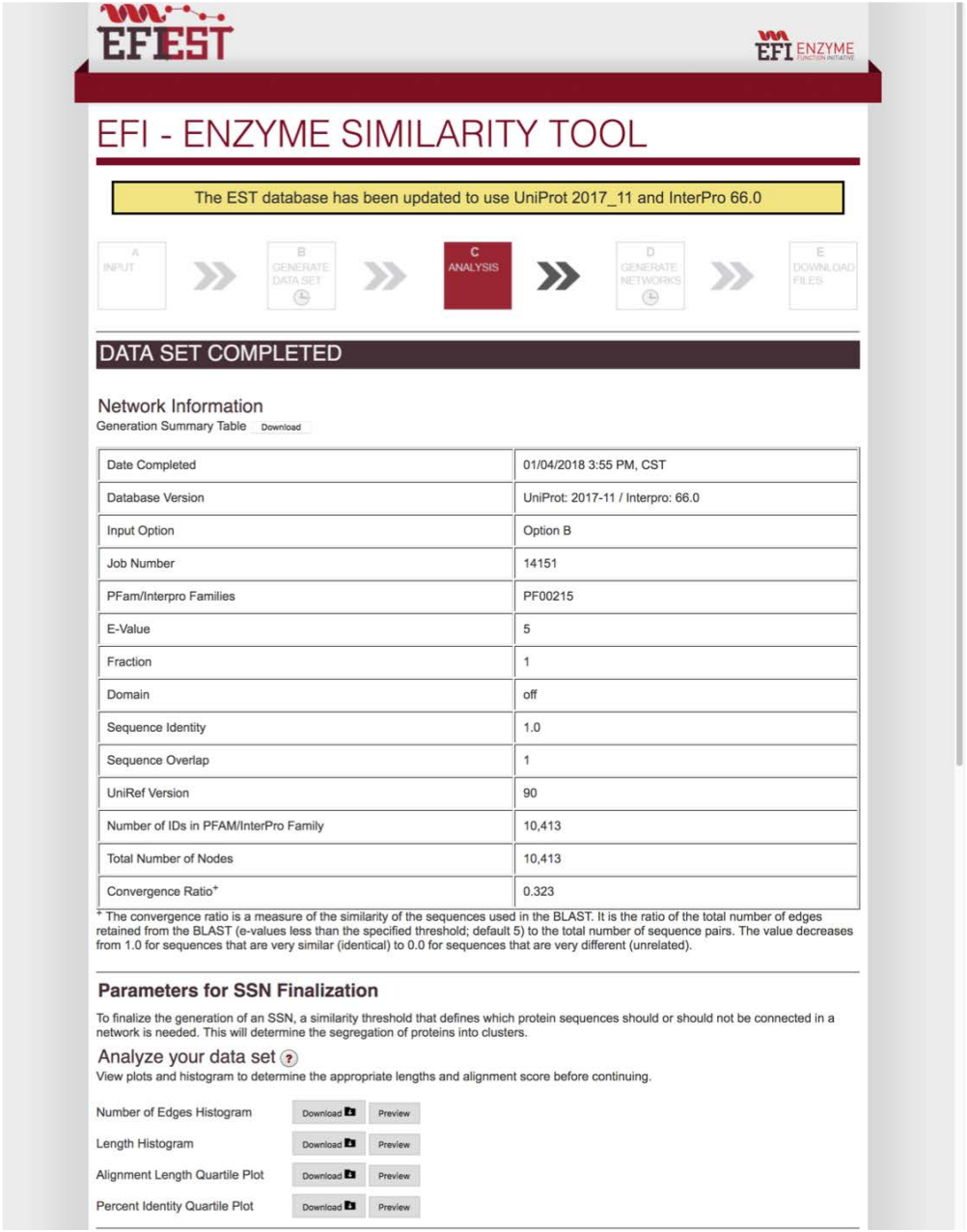
Option C: FASTA header with NCBI IDs

>**WP_012091266.1** hydroxyproline-2-epimerase [Ochrobactrum anthropi]
>**A6WXX7.1** RecName: Full=4-hydroxyproline 2-epimerase 2; Short=4Hyp 2-epimerase 2; Short=4HypE 2
>**ABS13831.1** proline racemase [Ochrobactrum anthropi ATCC 49188]
>**AIK43759.1** 4-hydroxyproline epimerase [Ochrobactrum anthropi]
MARHSFFCVDGHTCGNPVRLVAGGGPNLEGSTMMEKRAHFLREYDWIRTGLMFEPRGHDMMSG
SILYPPTRPDCDVAVLFIETSGCLPMCGHGTIGTVTMAIEQGLVTPKTPGKLNLDTPAGLVAIEYEQNG
QYVERVRLTNVPAFLYAEGLEVECPDLGNLKVDVAYGGNFYAIVEPQENYTDMEDYSALQLIAWSPILR
ERLNEKYKFQHPLLPDINRLSHILWTGKPKHPEAHARNAVFYGDKAIDRSPCGTGTSARMAQLAAKG
KLKPGDEFVHESIIGSLFHGRVERATEVVGQDRTLPAIIPSIAGWARMTGYNTIFIDDRDPFAHGFTVA

Option D: UniProt and/or NCBI IDs



Data Set Completed page



EFIEST

EFI ENZYME

EFI - ENZYME SIMILARITY TOOL

The EST database has been updated to use UniProt 2017_11 and InterPro 66.0

A INPUT » B GENERATE DATA SET » C ANALYSIS » D GENERATE NETWORKS » E DOWNLOAD FILES

DATA SET COMPLETED

Network Information

Generation Summary Table Download

Date Completed	01/04/2018 3:55 PM, CST
Database Version	UniProt: 2017-11 / Interpro: 66.0
Input Option	Option B
Job Number	14151
PFam/Interpro Families	PF00215
E-Value	5
Fraction	1
Domain	off
Sequence Identity	1.0
Sequence Overlap	1
UniRef Version	90
Number of IDs in PFAM/InterPro Family	10,413
Total Number of Nodes	10,413
Convergence Ratio+	0.323

+ The convergence ratio is a measure of the similarity of the sequences used in the BLAST. It is the ratio of the total number of edges retained from the BLAST (e-values less than the specified threshold; default 5) to the total number of sequence pairs. The value decreases from 1.0 for sequences that are very similar (identical) to 0.0 for sequences that are very different (unrelated).

Parameters for SSN Finalization

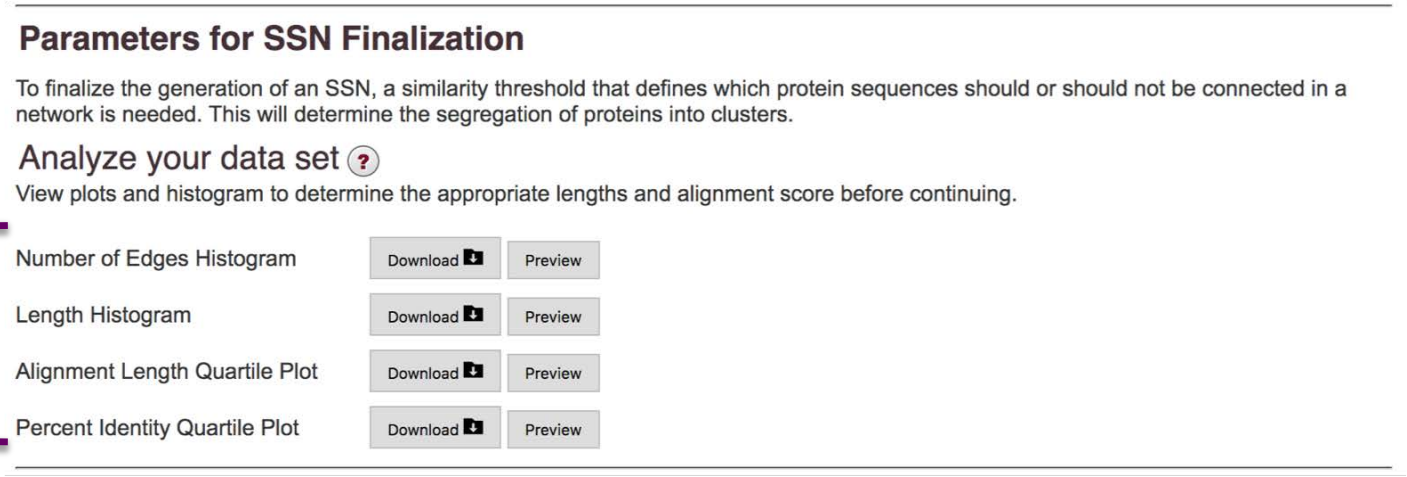
To finalize the generation of an SSN, a similarity threshold that defines which protein sequences should or should not be connected in a network is needed. This will determine the segregation of proteins into clusters.

Analyze your data set

View plots and histogram to determine the appropriate lengths and alignment score before continuing.

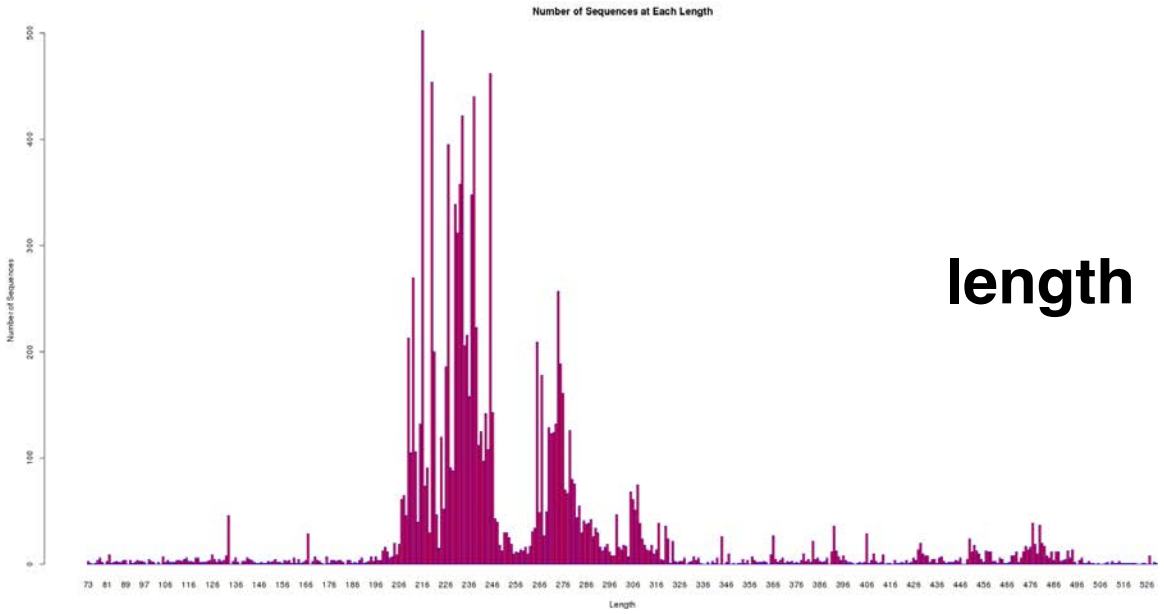
Number of Edges Histogram	Download	Preview
Length Histogram	Download	Preview
Alignment Length Quartile Plot	Download	Preview
Percent Identity Quartile Plot	Download	Preview

Data Set Completed page

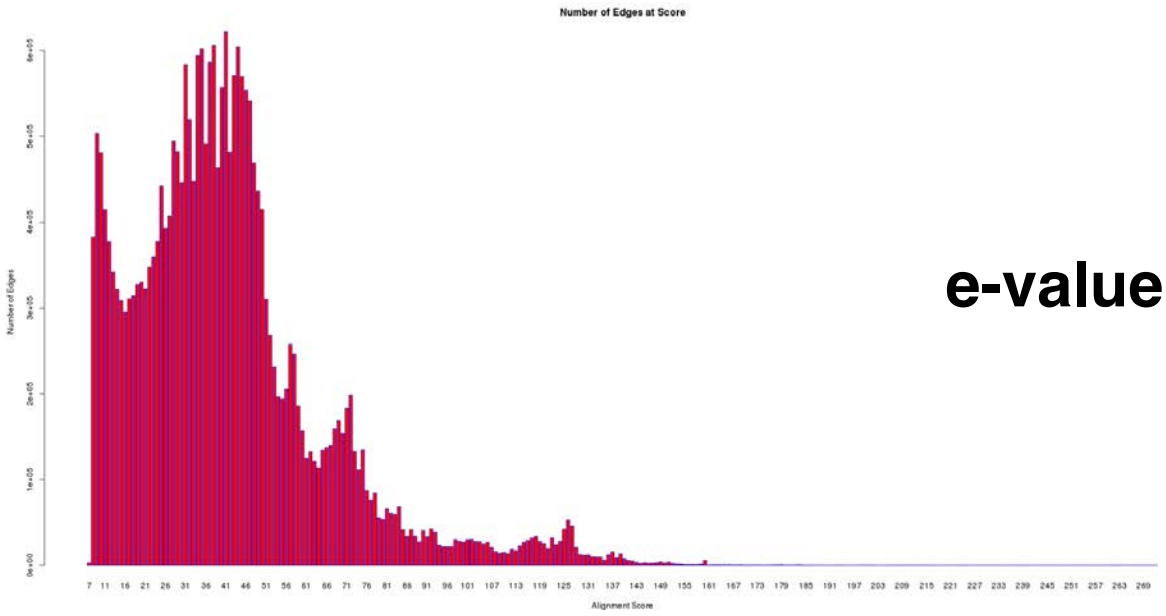


Statistical plots

Two histograms: length and e-value

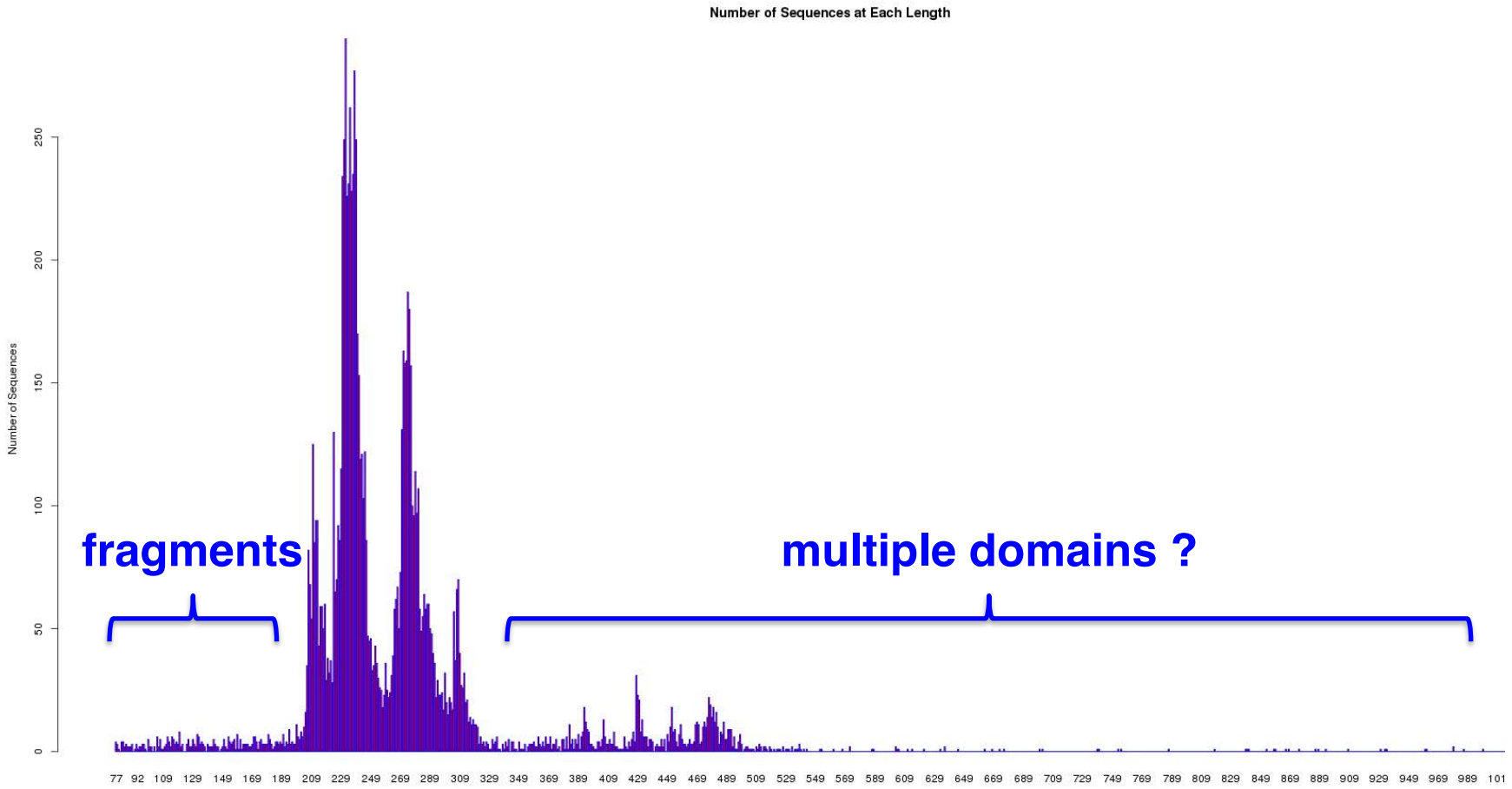


length

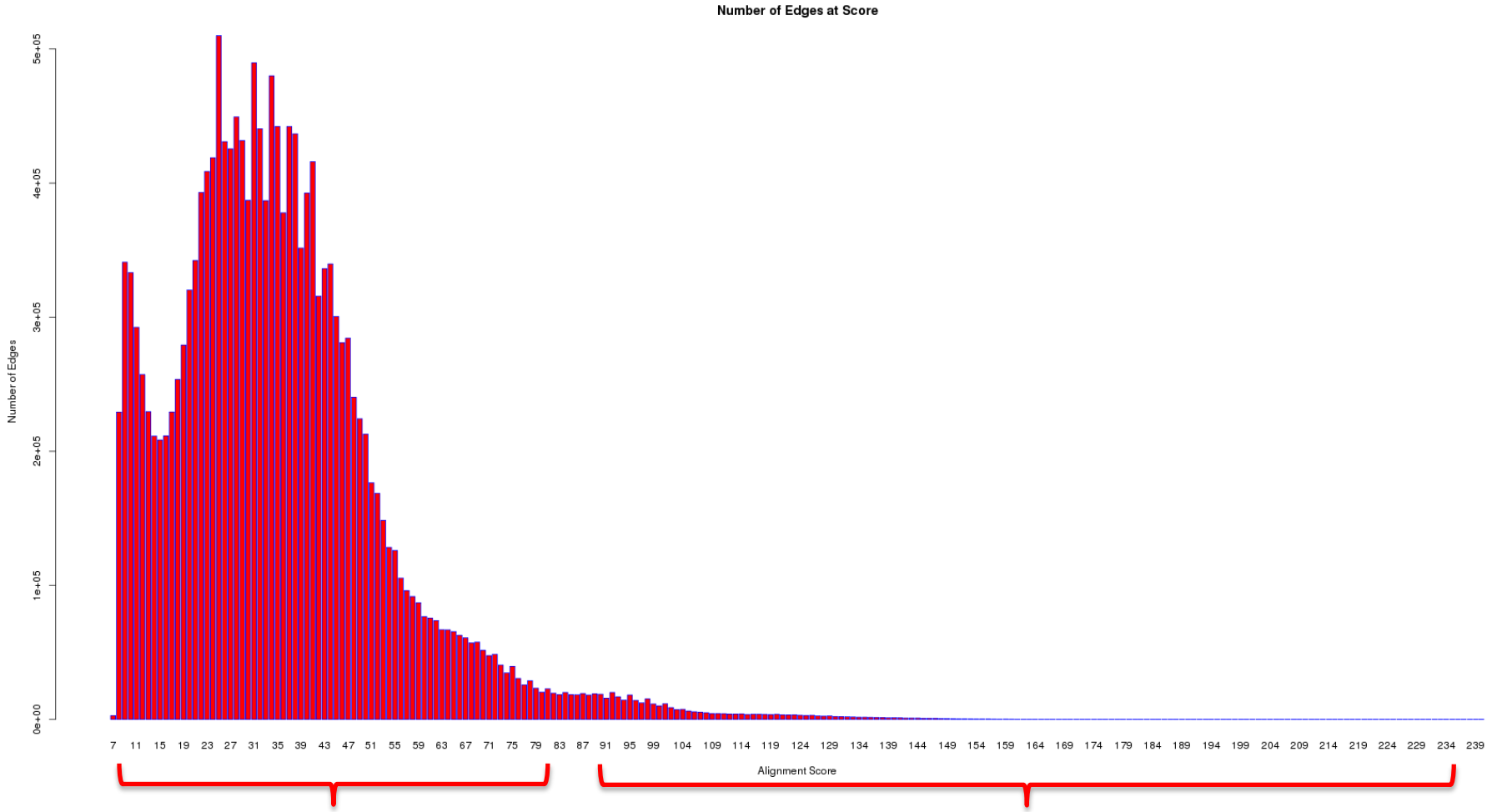


e-value

Length histogram: fragments and multiple domains



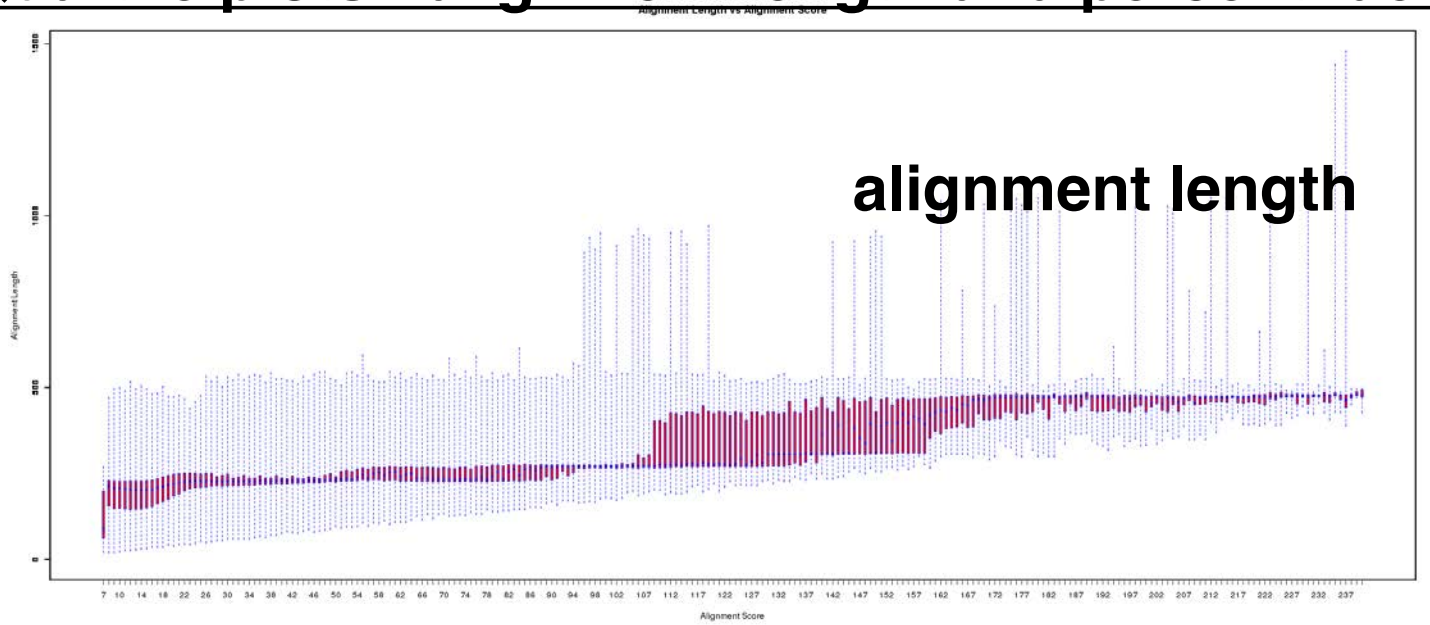
Edge histogram: sequence divergence

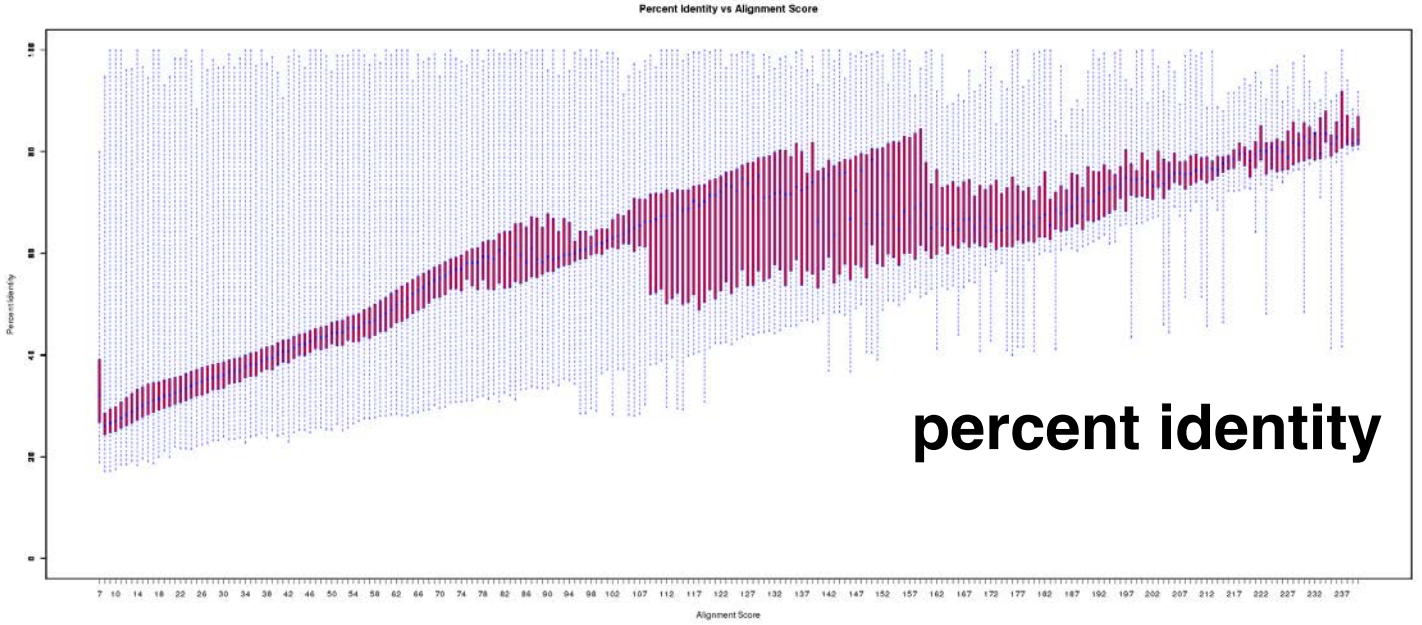


divergent connections between clusters

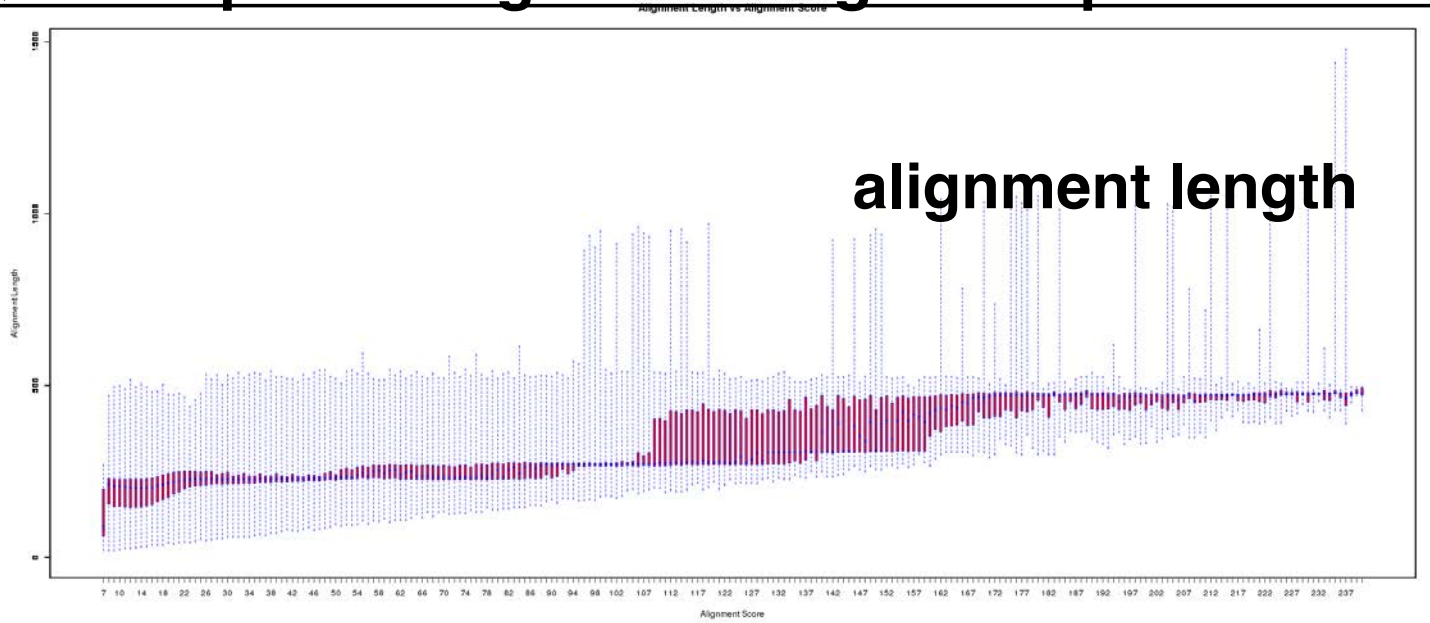
convergent, connections within clusters

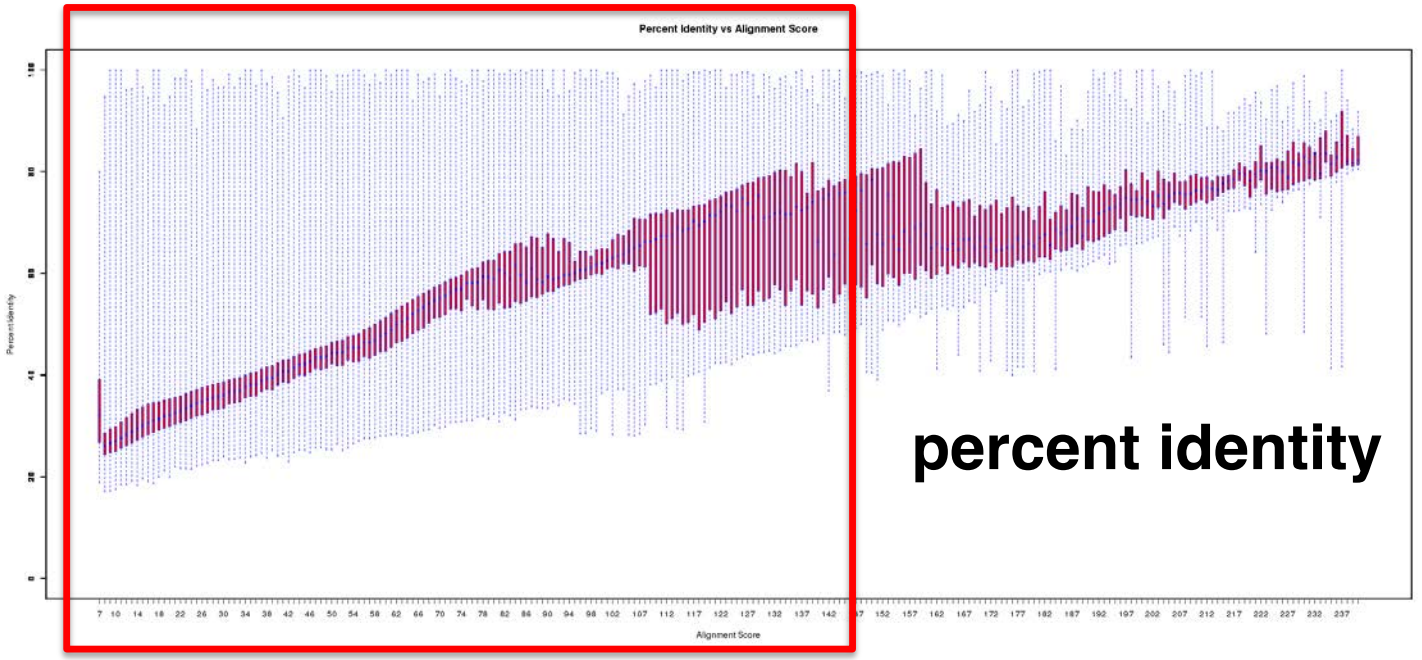
Quartile plots: alignment length and percent identity



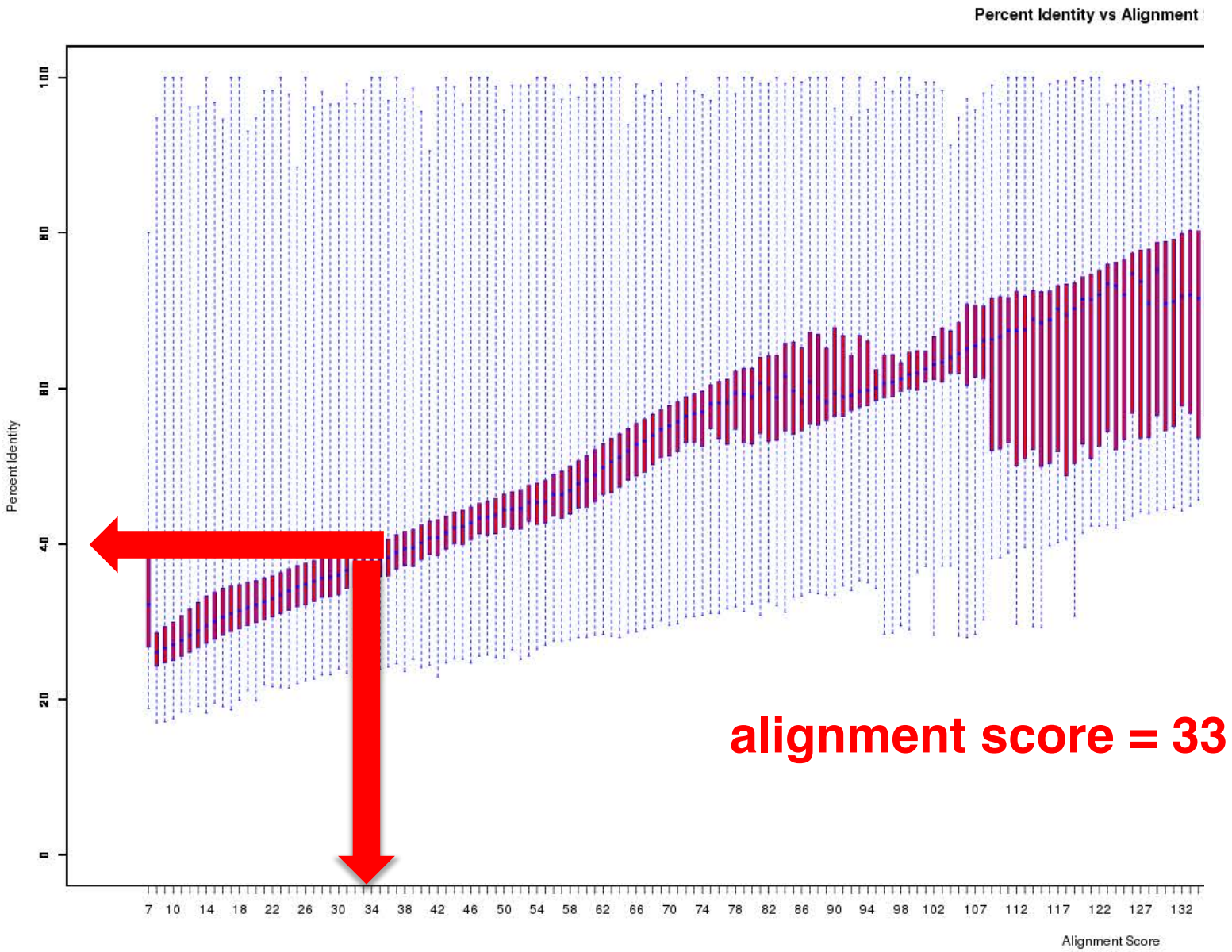


Quartile plots: alignment length and percent identity





Alignment score for initial SSN: ~40% identity

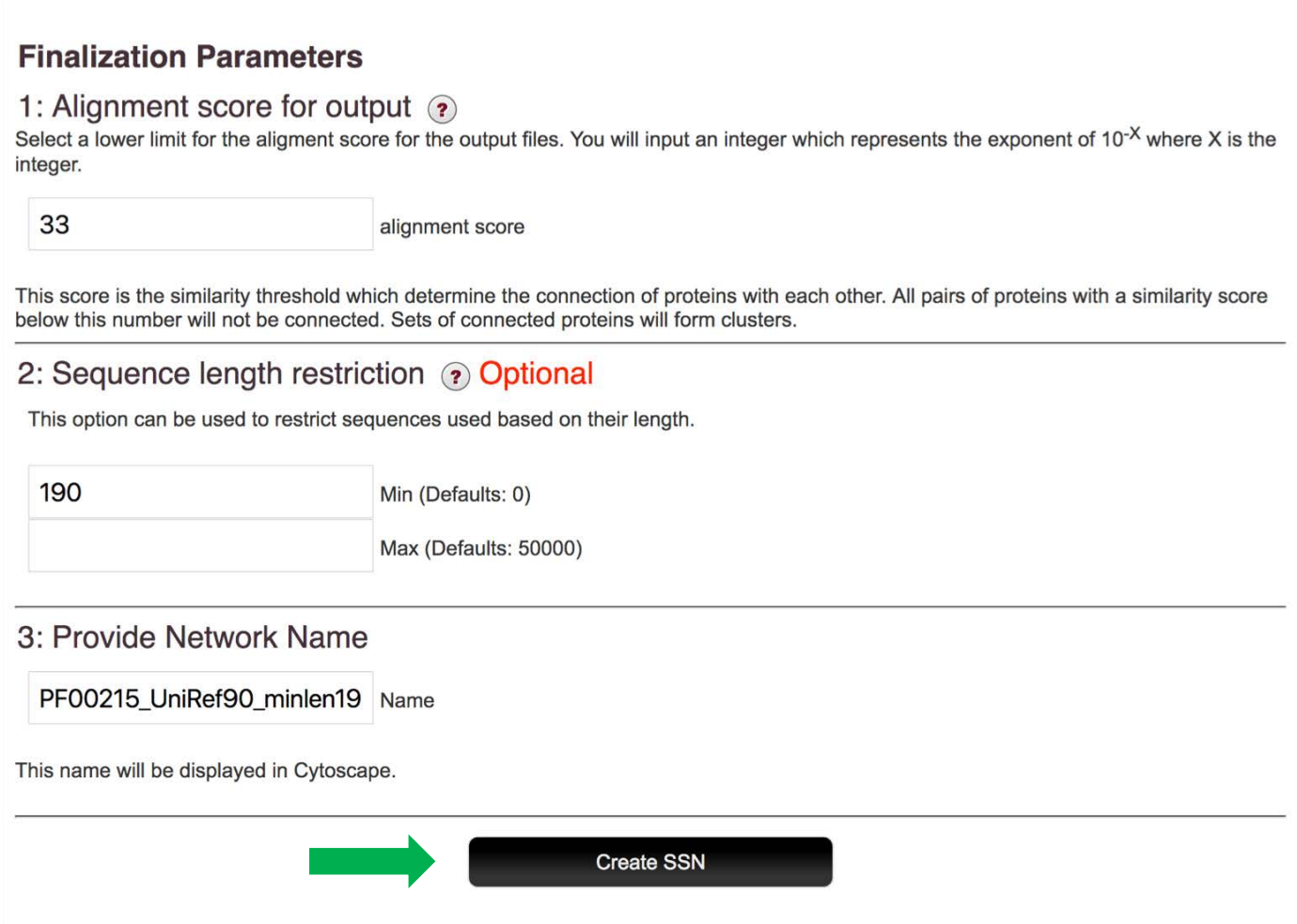


Analysis step: input alignment score and length ranges

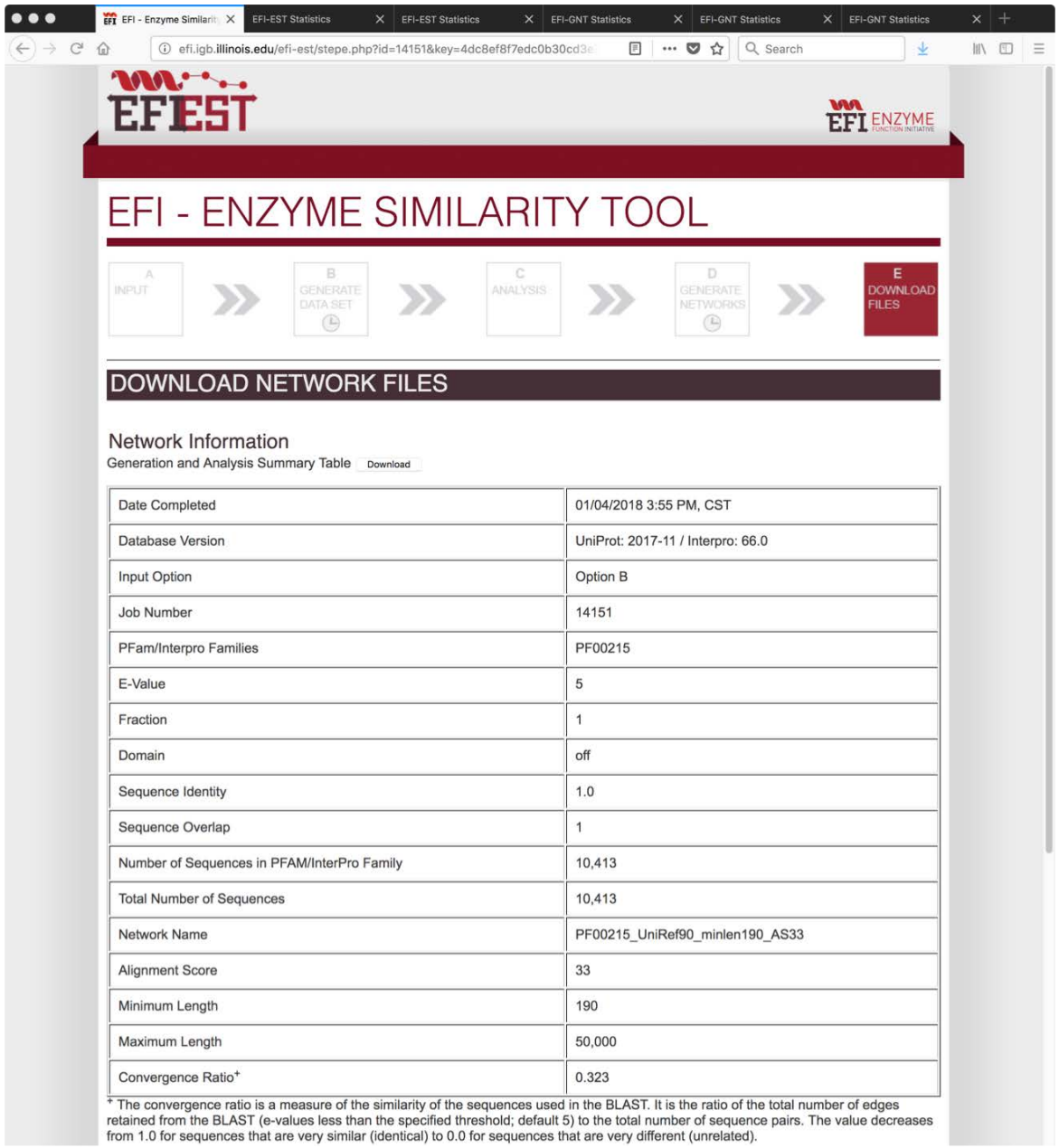
Alignment score

Length ranges

Network title



Download Network File page



EFI - Enzyme Similarity | EFI-EST Statistics | EFI-EST Statistics | EFI-GNT Statistics | EFI-GNT Statistics | EFI-GNT Statistics

efi.igb.illinois.edu/efi-est/stepe.php?id=14151&key=4dc8ef8f7edc0b30cd3e

EFIEST

EFI ENZYME FUNCTION INITIATIVE

EFI - ENZYME SIMILARITY TOOL

A INPUT → B GENERATE DATA SET → C ANALYSIS → D GENERATE NETWORKS → E DOWNLOAD FILES

DOWNLOAD NETWORK FILES

Network Information

Generation and Analysis Summary Table Download

Date Completed	01/04/2018 3:55 PM, CST
Database Version	UniProt: 2017-11 / Interpro: 66.0
Input Option	Option B
Job Number	14151
PFam/Interpro Families	PF00215
E-Value	5
Fraction	1
Domain	off
Sequence Identity	1.0
Sequence Overlap	1
Number of Sequences in PFAM/InterPro Family	10,413
Total Number of Sequences	10,413
Network Name	PF00215_UniRef90_minlen190_AS33
Alignment Score	33
Minimum Length	190
Maximum Length	50,000
Convergence Ratio+	0.323

+ The convergence ratio is a measure of the similarity of the sequences used in the BLAST. It is the ratio of the total number of edges retained from the BLAST (e-values less than the specified threshold; default 5) to the total number of sequence pairs. The value decreases from 1.0 for sequences that are very similar (identical) to 0.0 for sequences that are very different (unrelated).

Network statistics and downloads

Full network

	# Nodes	# Edges	File Size (MB)
Download Download ZIP	10,045	8,264,486	2,201 MB

Representative Node Networks

In representative node (RepNode) networks, each node in the network represents a collection of proteins grouped according to percent identity. For example, for a 75% identity RepNode network, all connected sequences that share 75% or more identity are grouped into a single node (meta node).

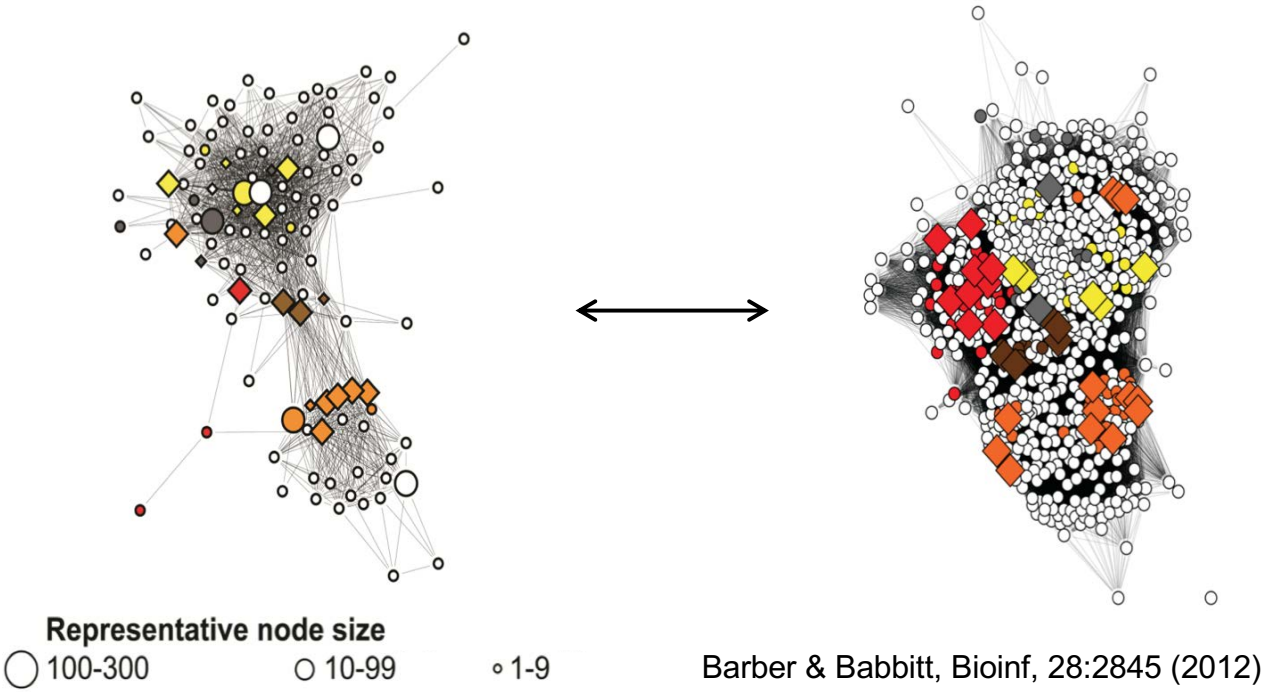
The cluster organization is not changed, and the clustering of sequences remains identical to the full network.

Sequences are collapsed together to reduce the overall number of nodes making for less complicated networks easier to load in Cytoscape.

Rep node networks

	% ID	# Nodes	# Edges	File Size (MB)
Download Download ZIP	100	10,045	8,264,486	2,214 MB
Download Download ZIP	95	10,004	8,187,043	2,194 MB
Download Download ZIP	90	9,888	7,993,153	2,143 MB
Download Download ZIP	85	8,987	6,479,231	1,743 MB
Download Download ZIP	80	8,073	5,120,215	1,383 MB
Download Download ZIP	75	7,265	4,105,975	1,114 MB
Download Download ZIP	70	6,608	3,320,182	905 MB
Download Download ZIP	65	5,943	2,596,128	712 MB
Download Download ZIP	60	5,231	1,877,192	521 MB
Download Download ZIP	55	4,459	1,234,862	349 MB
Download Download ZIP	50	3,694	731,963	214 MB
Download Download ZIP	45	2,951	384,690	120 MB
Download Download ZIP	40	2,269	185,689	66 MB

Representative node (rep node) networks



Barber & Babbitt, Bioinf, 28:2845 (2012)

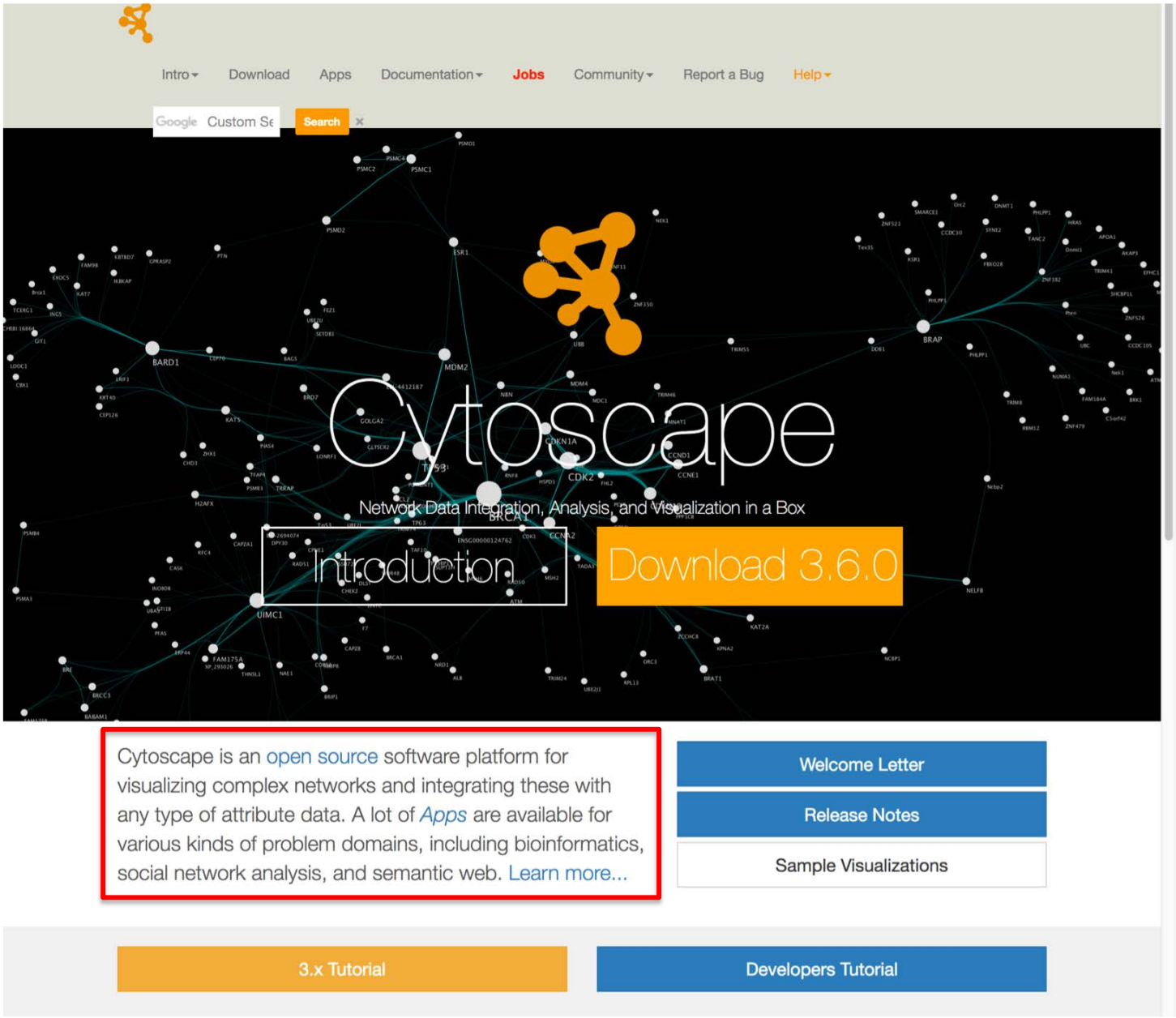
Visualization algorithms (Cytoscape) limit the number of edges (and, therefore, indirectly the number of nodes) that can be displayed: <~500K with 8GB RAM, <~1M with 16GB RAM; <~5M with 64GB RAM; <~10M with 128GB RAM

Solution: "bin" similar sequences into the same "metanode", e.g., 80%, 90%, or 100% sequence identity rep node networks

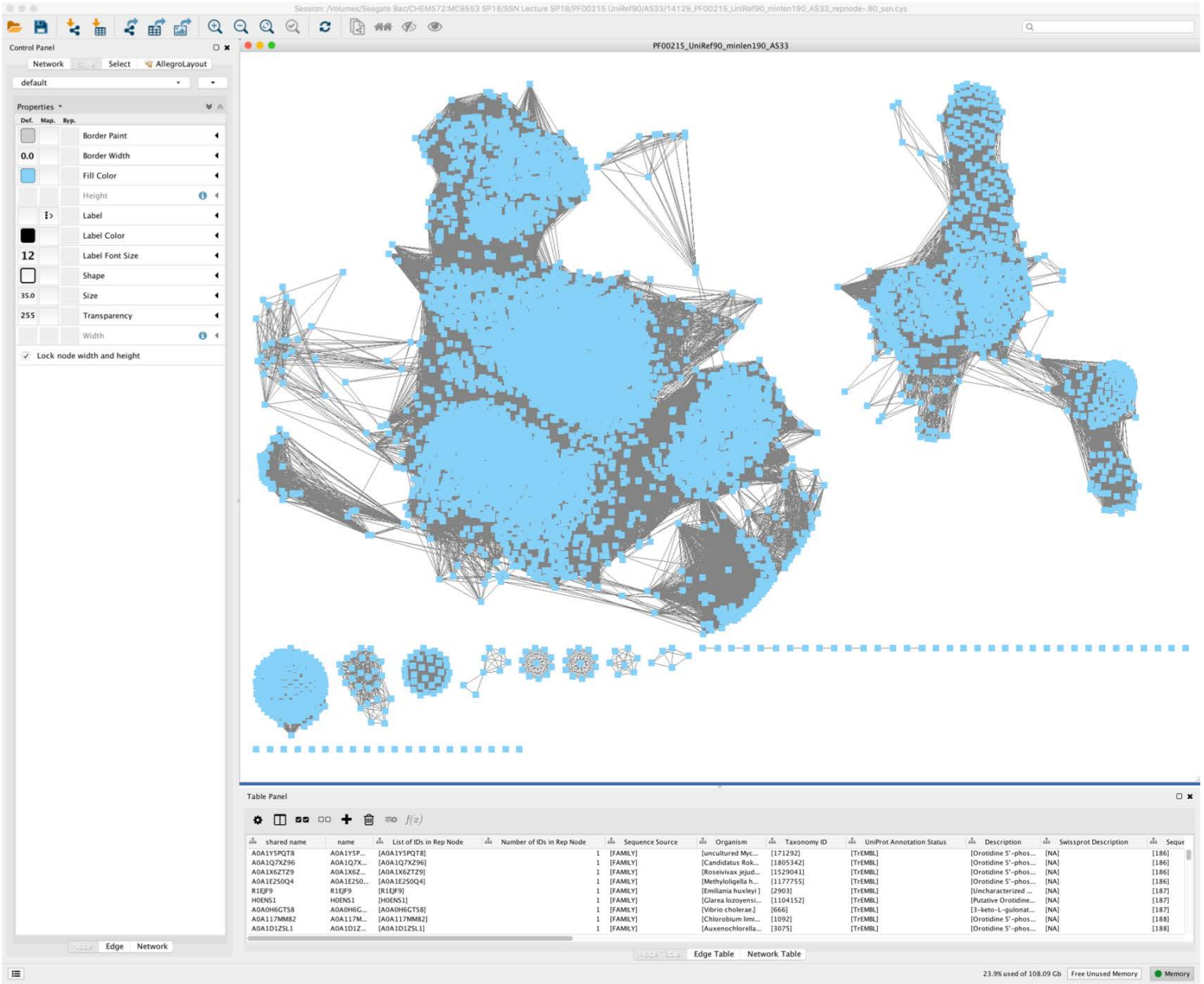
I don't get a commission! But …..

	Memory	Processor	RAM	Price
MacBook 12"	512GB	1.3GHz	8GB	$1,599
MacBook 12"	512GB	1.3GHz	16GB	$1,799
MacBook Pro 13"	512GB	3.1GHz	8GB	$1,999
MacBook Pro 13"	512GB	3.1GHz	16GB	$2,199
iMac 21"	1TB	3.4GHz	8GB	$1,499
iMac 21"	1TB	3.4GHz	16GB	$1,699
iMac 21"	1TB	3.4GHz	32GB	$2,099
iMac 27"	2TB	3.8 GHz	8GB	$2,299
iMac 27"	2TB	3.8 GHz	16GB	$2,499
iMac 27"	2TB	3.8 GHz	32GB	$2,899
iMac 27"	2TB	3.8 GHz	64GB	$3,699
iMac Pro	1TB	3.2GHz	32GB	$4,999
iMac Pro	1TB	3.2GHz	64GB	$5,799
iMac Pro	1TB	3.2GHz	128GB	$7,399

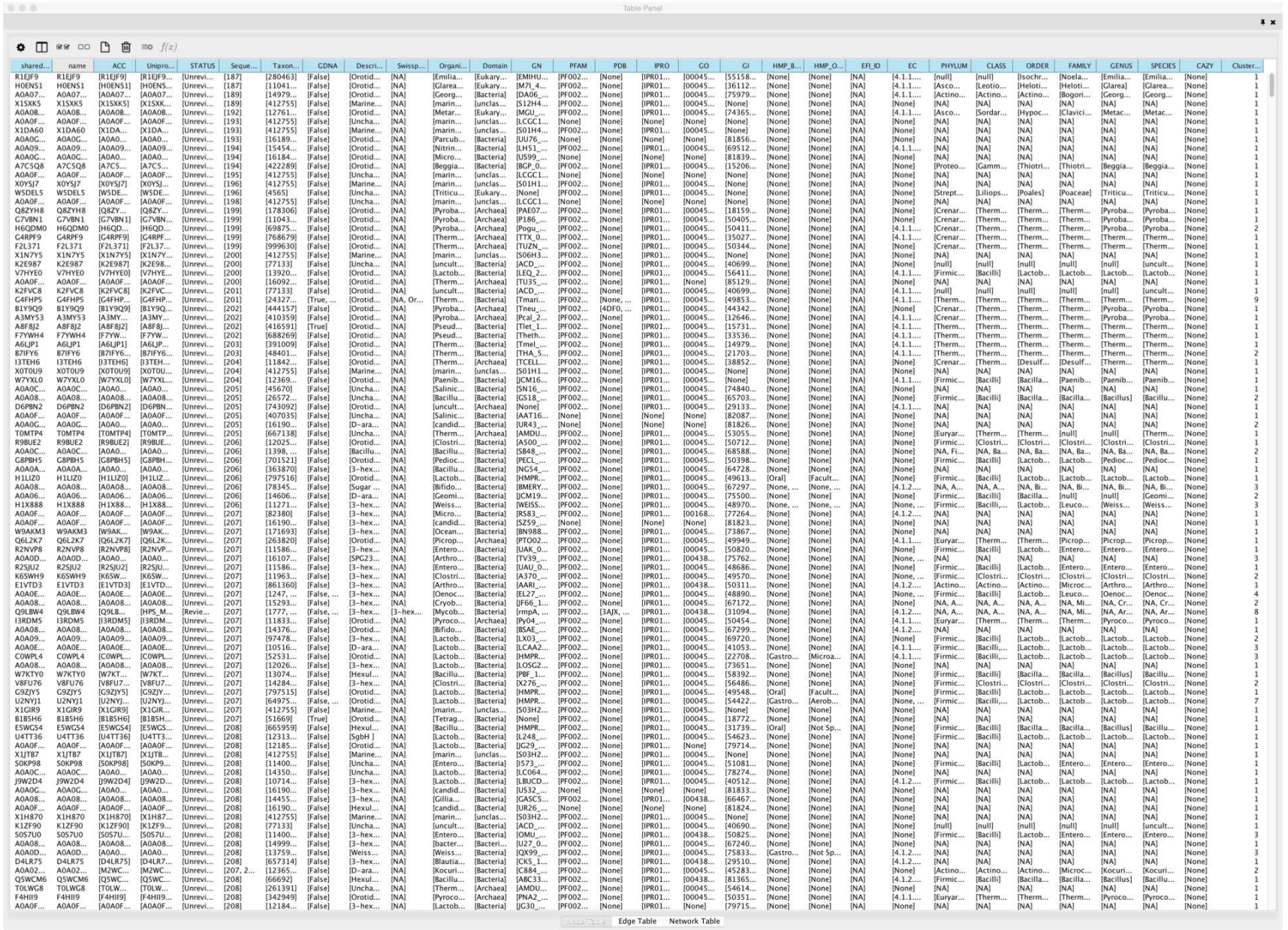
Cytoscape: http://cytoscape.org/



Cytoscape 3.3.0



Data panel: node attributes



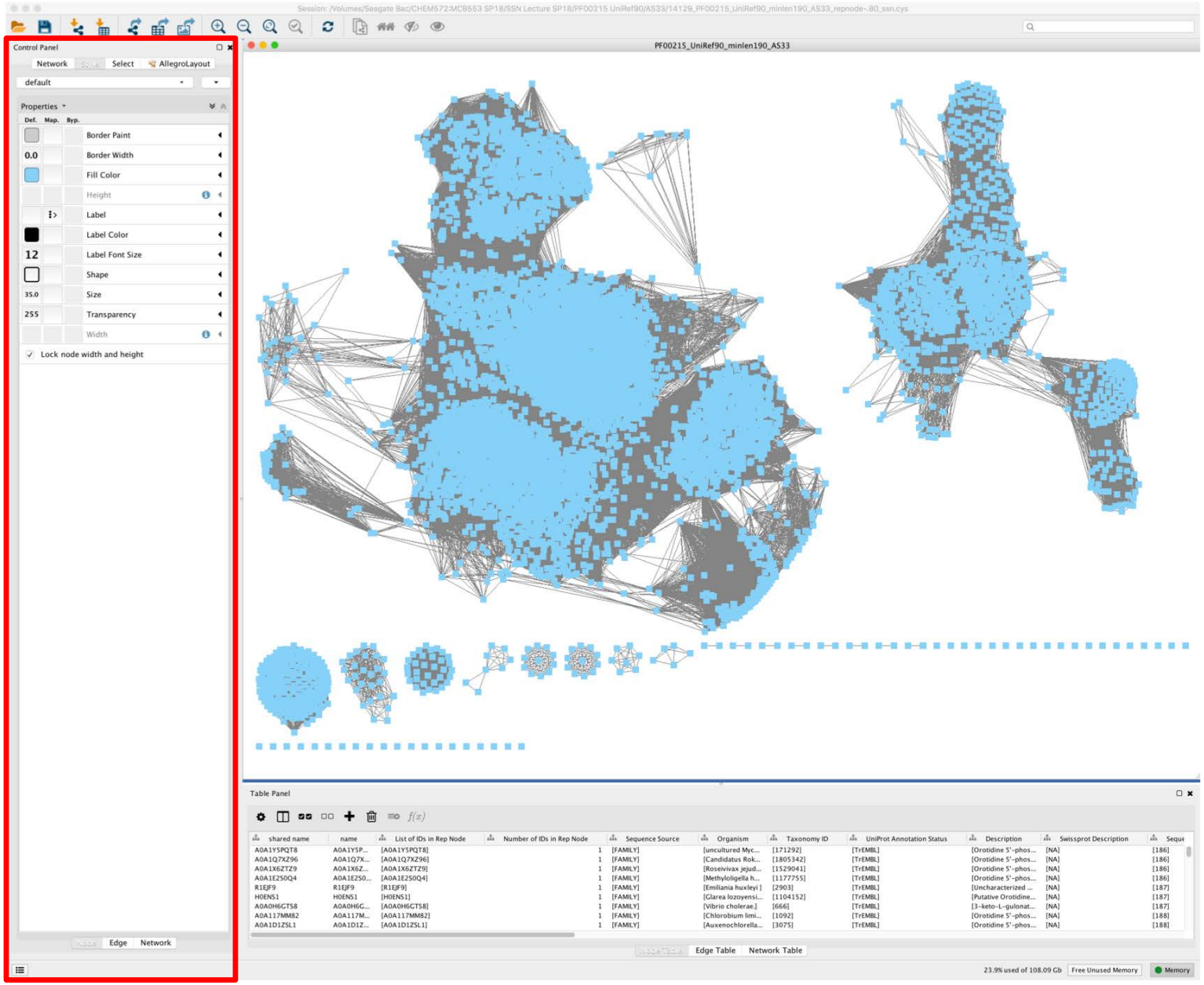
Node attributes: part 1

shared name – UniProt accession ID
name – UniProt accession ID
Organism – Organism name
Taxonomy ID – NCBI Taxnomy ID
UniProt Accession Status – TrEMBL or SwissProt
Description – UniProt description
Swissprot Description – Swiss Prot description
Sequence Length – number of residues
Gene Name – Gene name
NCBI IDs – RefSeq IDs, UniProt/Swiss-Prot IDs, GenBank IDs, PDB IDs, "retired" GI numbers
Superkingdom – Phylogenetic superkingdom of the organism (from UniProt)
Kingdom – Phylogenetic kingdom of the organism
Phylum – Phylogenetic phylum of the organism
Class – Phylogenetic class of the organism
Order – Phylogenetic order of the organism
Family – Phylogenetic family of the organism
Genus – Phylogenetic genus of the organism
Species – Phylogenetic species of the organism

Node attributes: part 2

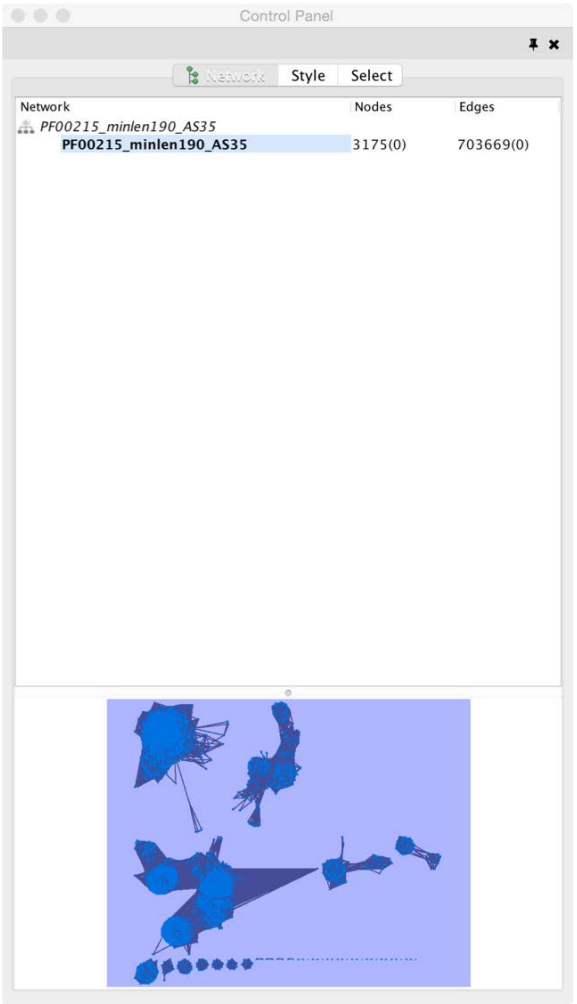
EC – EC (Enzyme Commisssion) number
PFAM – Pfam family number
IPRO – InterPro domain/family number
BRENDA ID – BRENDA ID: https://www.brenda-enzymes.org/
CAZY – Carbohydrate-Active enZYmes: http://www.cazy.org/
GO Term – Gene ontology (from UniProt)
KEGG ID – KEGG ID: http://www.genome.jp/kegg/
PATRIC ID – Genome annotation and assembly: https://www.patricbrc.org/
STRING ID – Protein-protein interactions: https://string-db.org/
HMP Body Site – Human Microbiome Project organisms
HMP Oxygen – Human Microbiome Project organisms
P01 gDNA - availability of gDNA in P01GM118303 Protein Core
UniRef90 Cluster IDs – UniProt IDs for seed sequences in UniRef90 clusters
UniRef90 Cluster Size – number of sequences/IDs in UniRef90 clusters
Sequence Source – Options B/C/D: USER or FAMILY
Query IDs – Option C/D: UniProt IDs in FASTA headers or input list
Other IDs – Option C: FASTA headers with unrecognized accession IDs
List of IDs in Repnode – UniProt IDs in repnode metanodes
Number of IDs in Repnode – number of sequences in repnode metanodes

Cytoscape 3.3.0: control panels

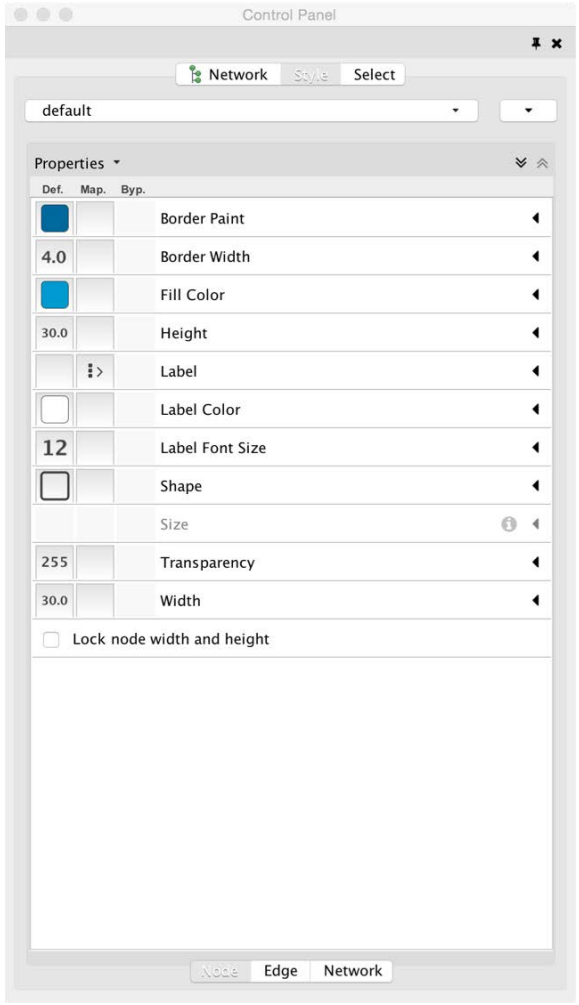


Cytoscape control panels: manipulating networks

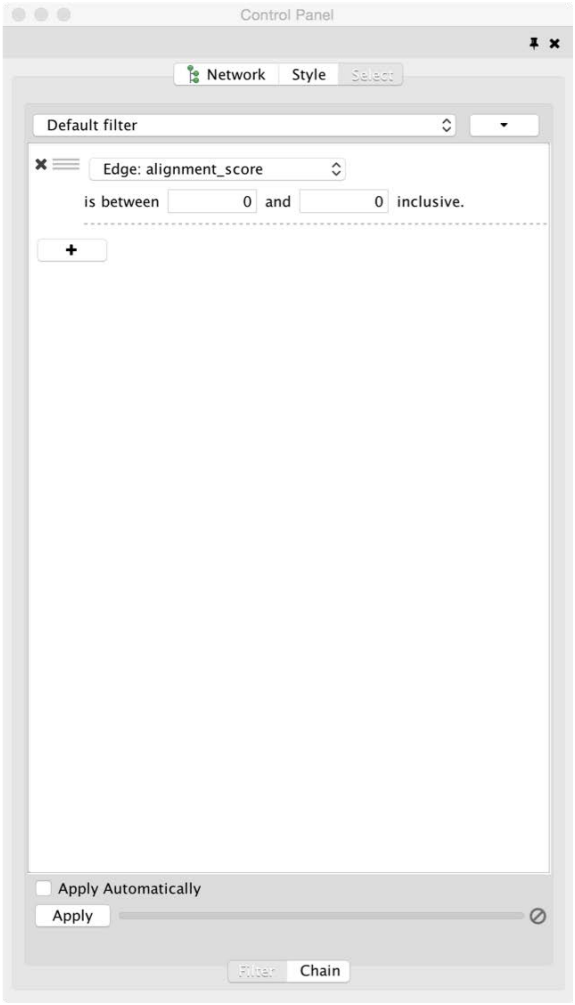
Network



Style



Select



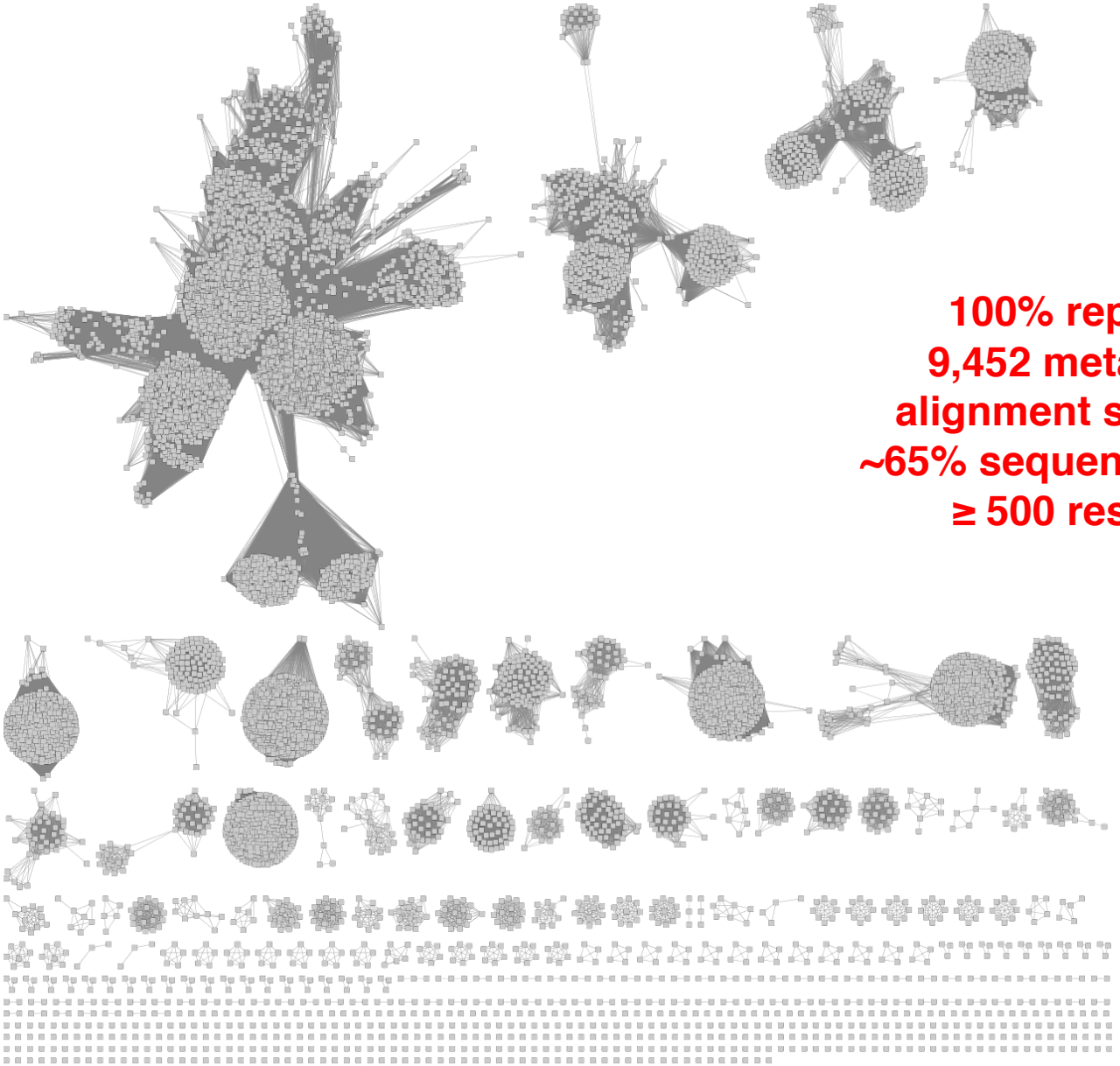
Some ”bioinorganic” enzyme superfamilies

Name	Pfam Family	InterPro Family	Family Size
Radical SAM	PF04055	IPR007197	360,840
P450	PF00067	IPR001128	211,400
Ferritin-like		IPR009078	168,246
Molybdopterin oxidoreductase	PF00384	IPR006656	94,758
Oxoglutarate Fe-independent Dioxygenase	PF13640/PF03171	IPR005123	84,800
Isopenicillin N synthase		IPR027443	50,698
TauD/TfdA-like	PF02668	IPR003819	35,996
Phytanoyl-CoA dioxygenase	PF05721	IPR008775	34,121
Ribonucleotide reductase small subunit family	PF00268	IPR000358	23,764
Oxoglutarate AlkB-like	PF13532	IPR027450	18,414
Glycyl radical enzyme	PF02901	IPR004184	14,229
FldB/FldC dehydratase	PF06050	IPR010327	9,079
2OG-Fe dioxygenase	PF10014	IPR018724	2,765
2OGFeDO oxygenase domain	PF12851	IPR024779	930

Some ”bioinorganic” enzyme superfamilies

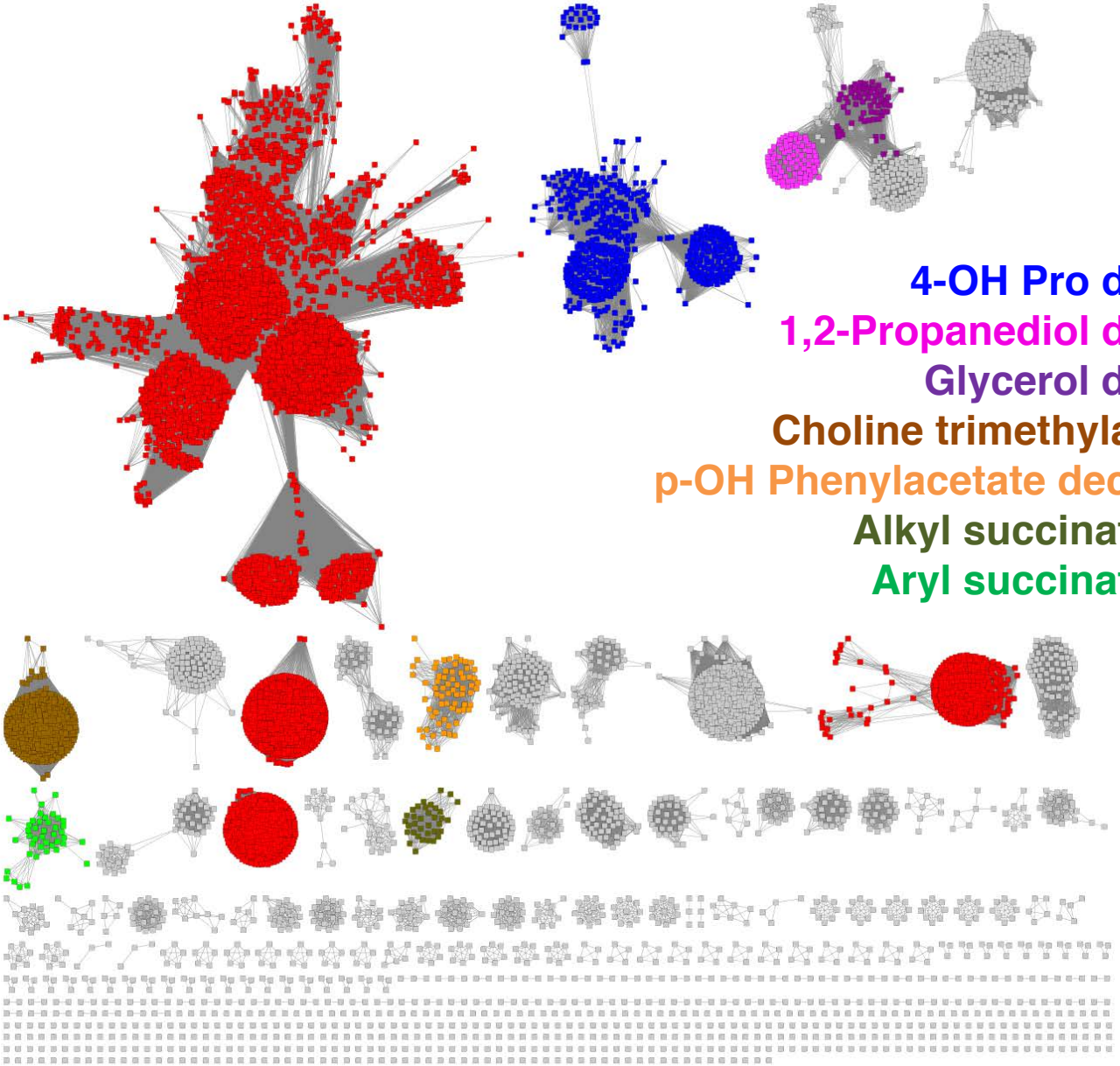
Name	Pfam Family	InterPro Family	Family Size
Radical SAM	PF04055	IPR007197	360,840
P450	PF00067	IPR001128	211,400
Ferritin-like		IPR009078	168,246
Molybdopterin oxidoreductase	PF00384	IPR006656	94,758
Oxoglutarate Fe-independent Dioxygenase	PF13640/PF03171	IPR005123	84,800
Isopenicillin N synthase		IPR027443	50,698
TauD/TfdA-like	PF02668	IPR003819	35,996
Phytanoyl-CoA dioxygenase	PF05721	IPR008775	34,121
Ribonucleotide reductase small subunit family	PF00268	IPR000358	23,764
Oxoglutarate AlkB-like	PF13532	IPR027450	18,414
Glycyl radical enzyme	**PF02901**	**IPR004184**	**14,229**
FldB/FldC dehydratase	PF06050	IPR010327	9,079
2OG-Fe dioxygenase	PF10014	IPR018724	2,765
2OGFeDO oxygenase domain	PF12851	IPR024779	930

Glycyl radical enzyme (GRE) superfamily: 14,229 sequence



100% rep node
9,452 metanodes
alignment score 300
~65% sequence identity
≥ 500 residues

Clusters with known functions

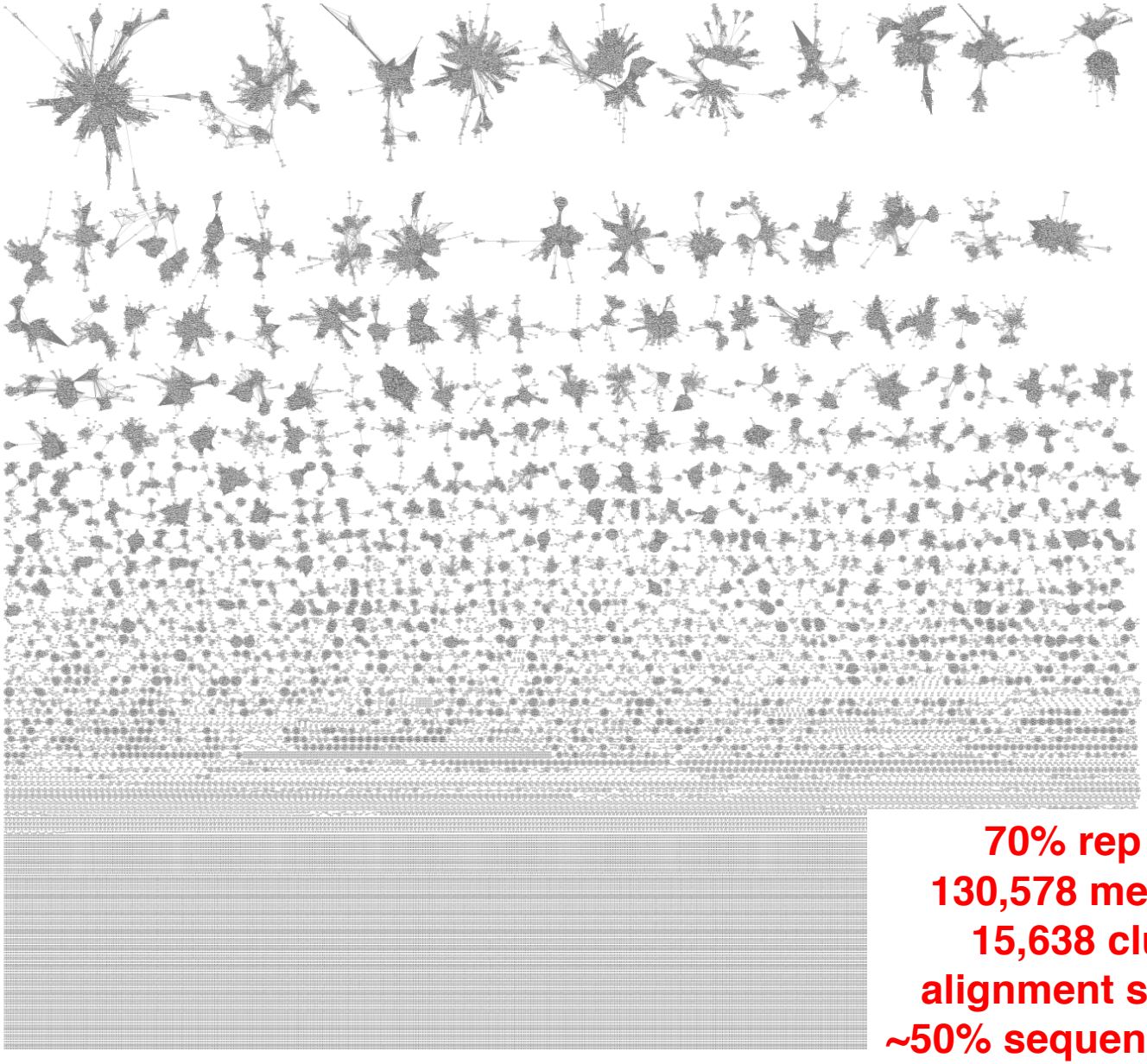


PFL
4-OH Pro dehydratase
1,2-Propanediol dehydratase
Glycerol dehydratase
Choline trimethylamine lyase
p-OH Phenylacetate decarboxylase
Alkyl succinate synthase
Aryl succinate synthase

Some ”bioinorganic” enzyme superfamilies

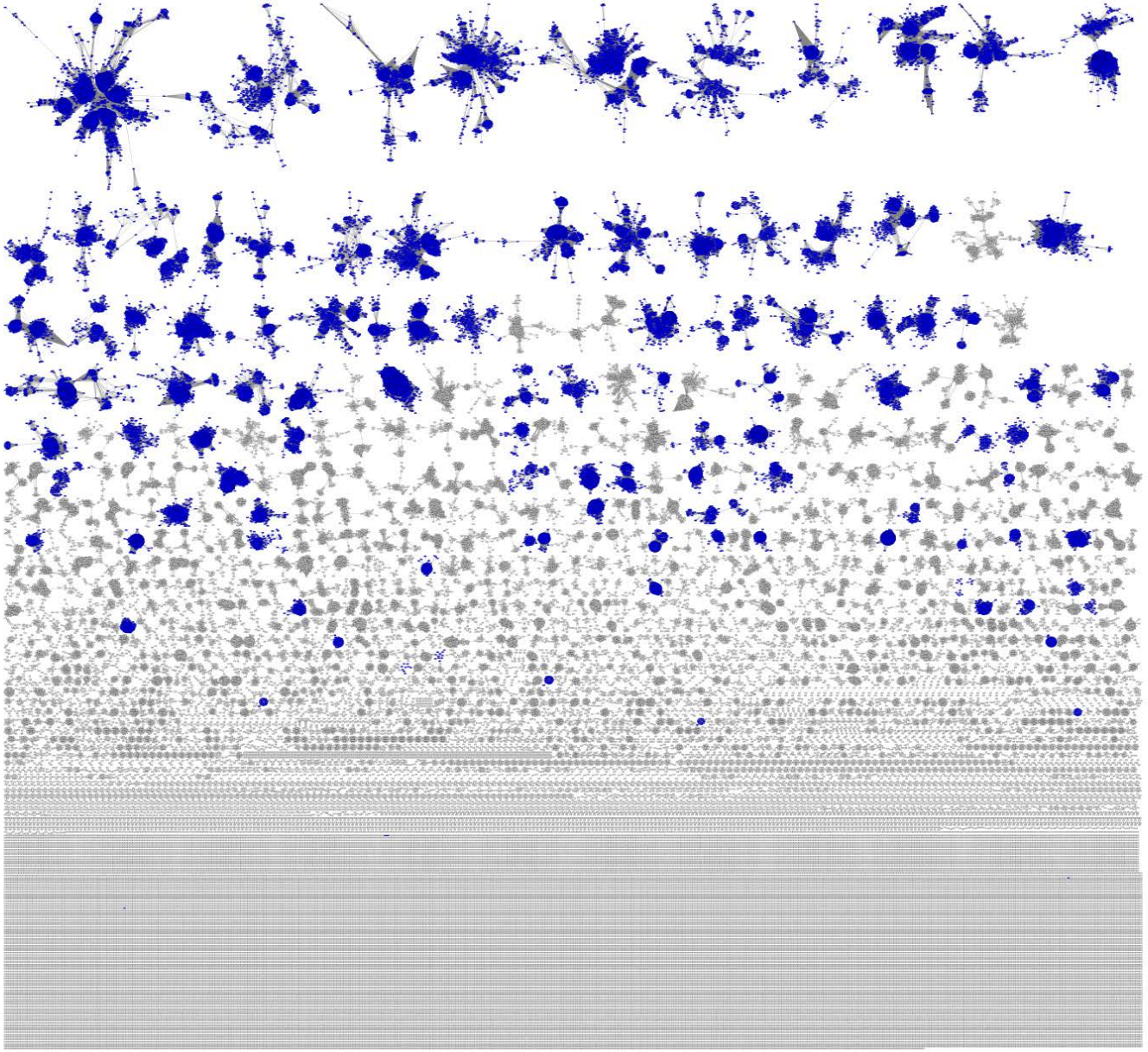
Name	Pfam Family	InterPro Family	Family Size
Radical SAM	**PF04055**	**IPR007197**	**360,840**
P450	PF00067	IPR001128	211,400
Ferritin-like		IPR009078	168,246
Molybdopterin oxidoreductase	PF00384	IPR006656	94,758
Oxoglutarate Fe-independent Dioxygenase	PF13640/PF03171	IPR005123	84,800
Isopenicillin N synthase		IPR027443	50,698
TauD/TfdA-like	PF02668	IPR003819	35,996
Phytanoyl-CoA dioxygenase	PF05721	IPR008775	34,121
Ribonucleotide reductase small subunit family	PF00268	IPR000358	23,764
Oxoglutarate AlkB-like	PF13532	IPR027450	18,414
Glycyl radical enzyme	PF02901	IPR004184	14,229
FldB/FldC dehydratase	PF06050	IPR010327	9,079
2OG-Fe dioxygenase	PF10014	IPR018724	2,765
2OGFeDO oxygenase domain	PF12851	IPR024779	930

Radical SAM superfamily: 360,891 sequences



70% rep node
130,578 metanodes
15,638 clusters
alignment score 120
~50% sequence identity

Clusters with SwissProt annotations



Squire Booker on radical SAM "unknown function space"

"These enzymes catalyze a dizzying array of novel transformations, and their discovery has led to a revival of mechanistic enzymology. Indeed, many beginning professors who are interested in chemical biology (i.e. natural product biosynthesis) or mechanistic enzymology find themselves working in the field of radical SAM enzymology. The ability to mine the large space of unknown function within the superfamily would be crucial to uncovering these new reactions and seeding new projects for young professors."

Today

1. Target selection using sequence similarity networks: EFI-EST web tool

2. **Pathway discovery using genome neighborhood networks: EFI-GNT web tool**

3. Example: using SSNs and GNNs to discover catabolic pathways for D-apiose

4. Target prioritization using chemically guided functional profiling: CGFP-ShortBRED beta tool

Use genome context to inform annotation for bacterial and fungal enzymes

High sequence similarity (≥ 70%) → **Enzyme function**

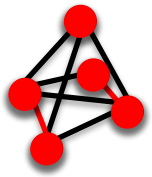
Low sequence similarity (≤ 70%) → **???????**

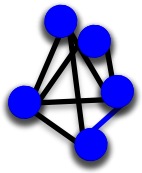
Low sequence similarity (≤ 70%) + **Genome context** → **Enzyme function**

Genome context in Genome Neighborhood Networks (GNNs)

Input: SSN families from EFI-EST

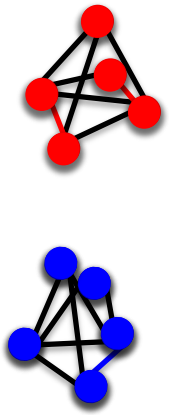
Query families



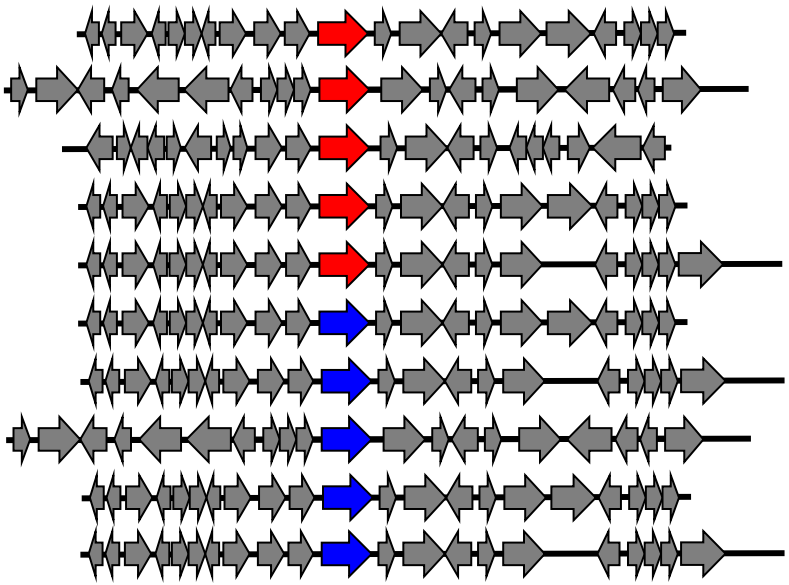


GNN: bacterial and fungal proteins in gene clusters

Query families

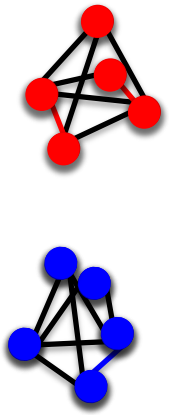


Genome neighbors

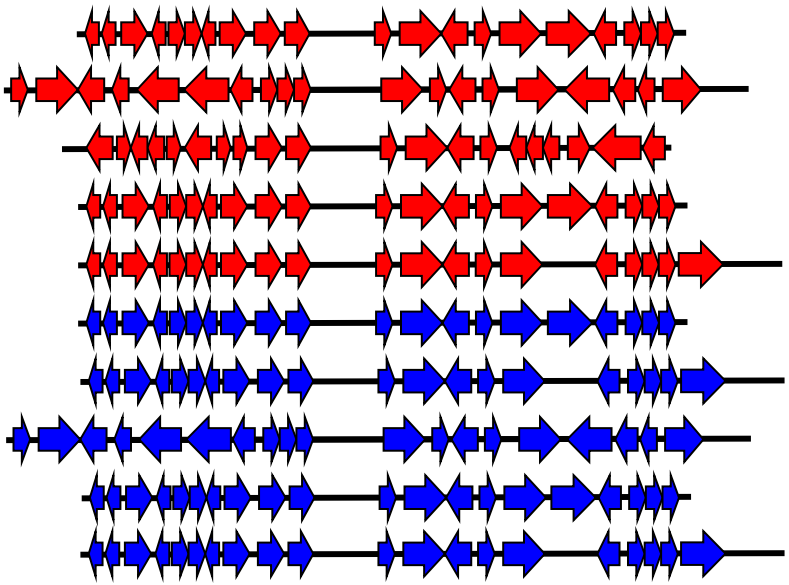


GNN: collect neighbors (±10)

Query families

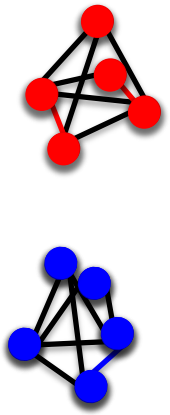


Genome neighbors

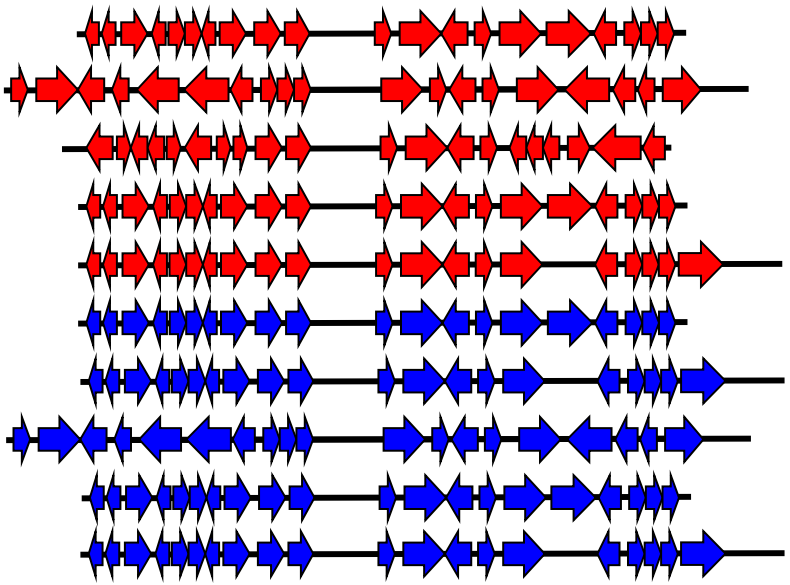


First GNN: Pfam families found by SSN clusters

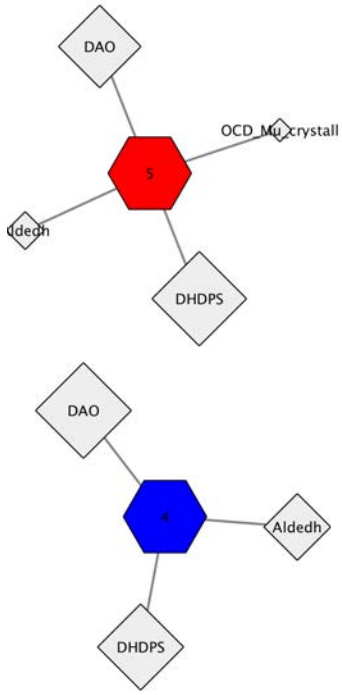
Query families



Genome neighbors

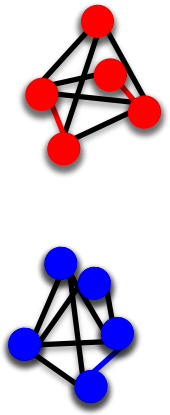


SSN cluster neighbors

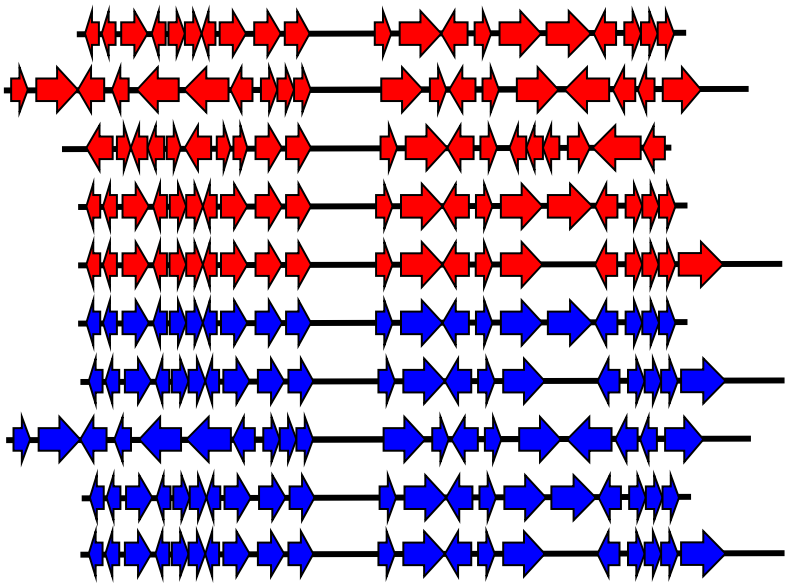


First GNN: "read" pathways from GNN clusters

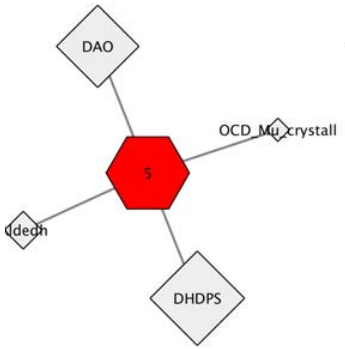
Query families

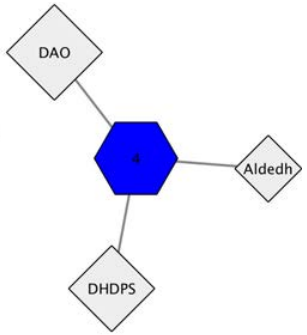


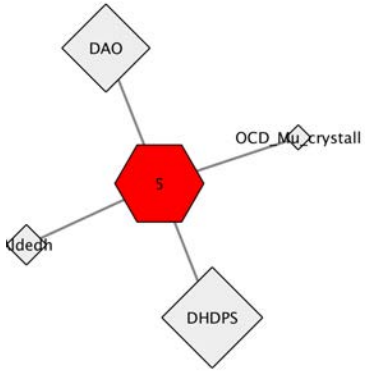
Genome neighbors

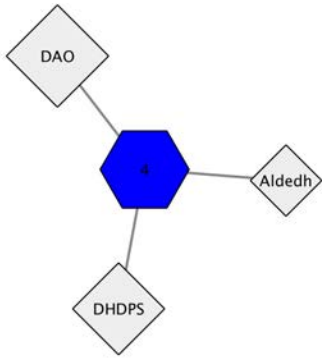


SSN cluster neighbors



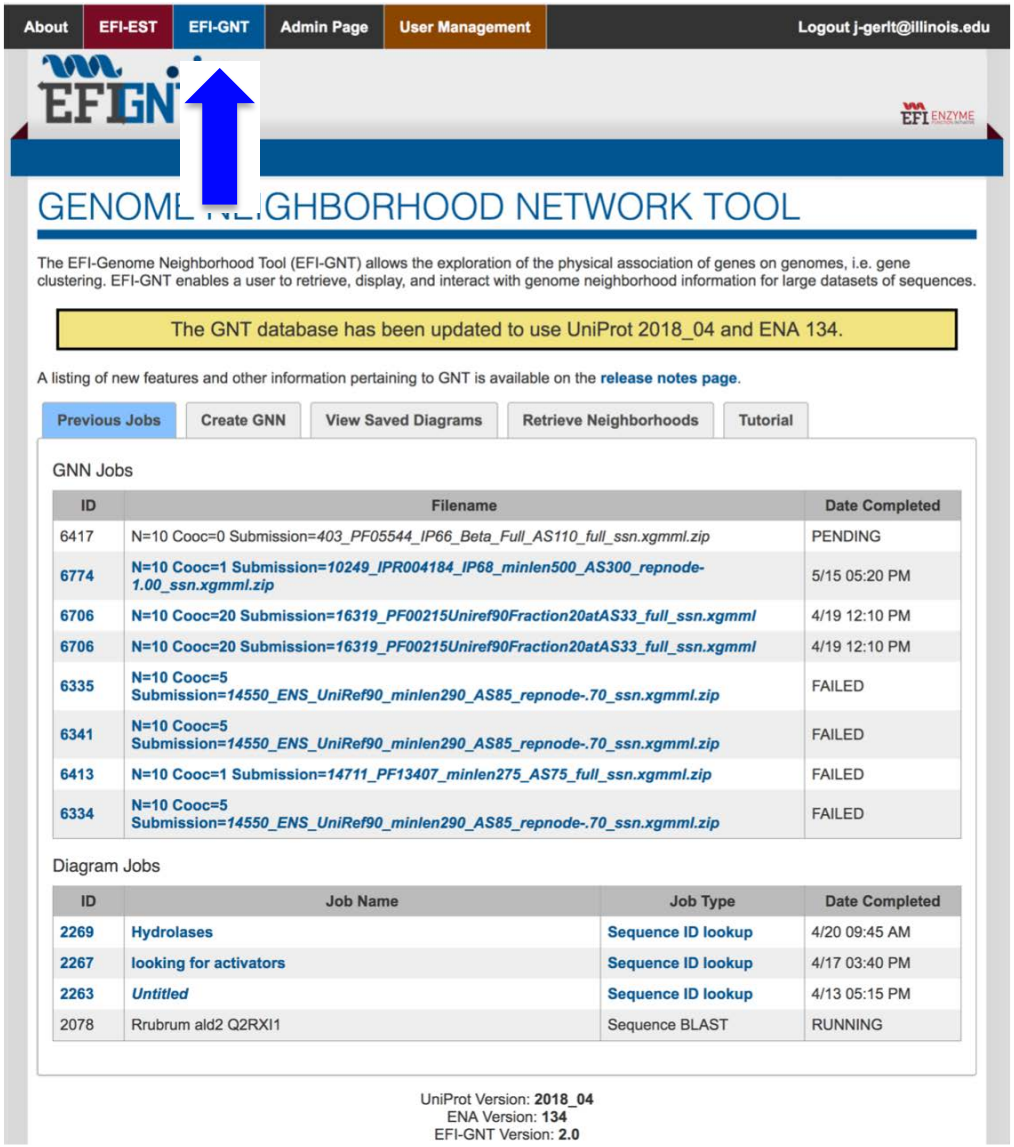






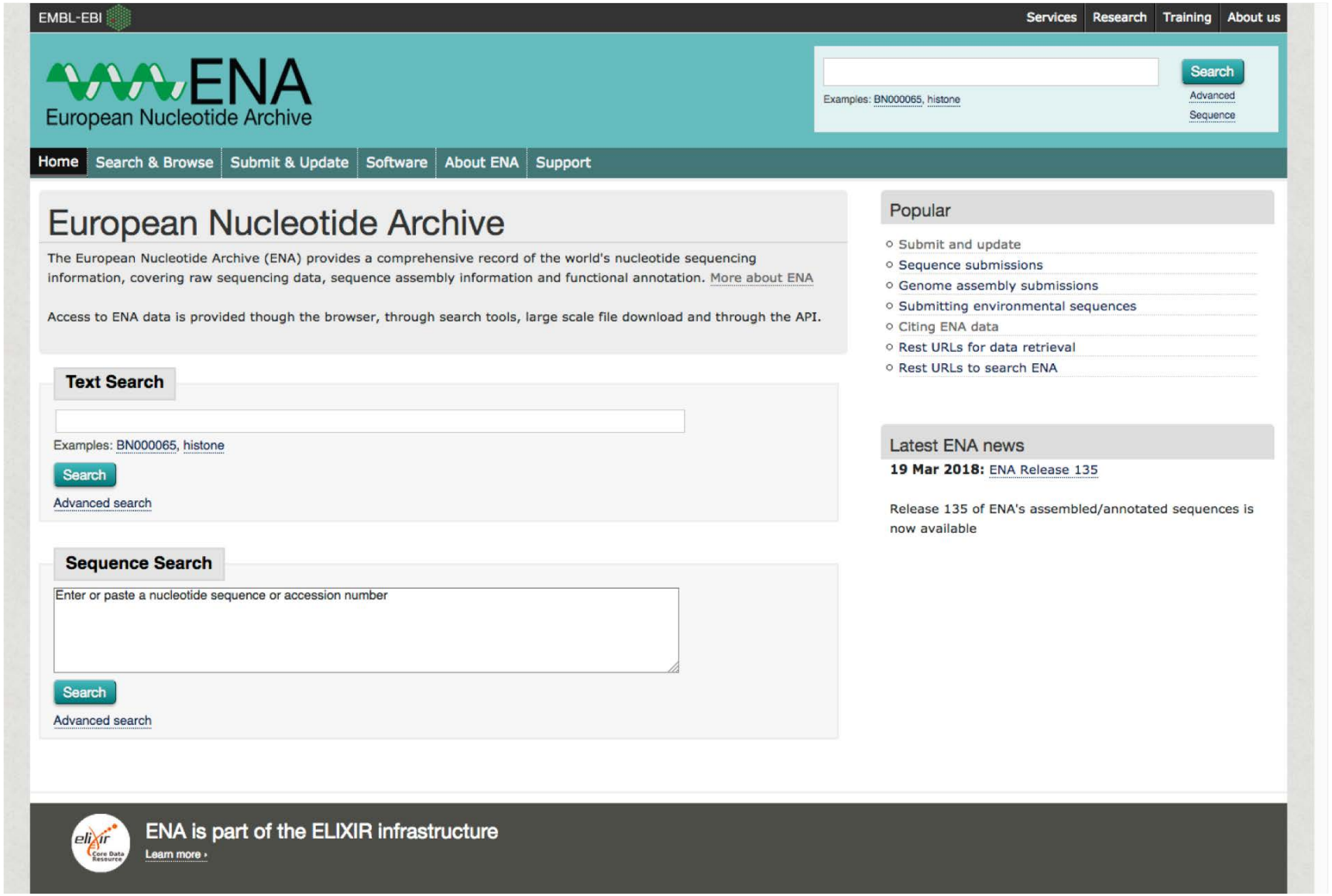
Deduce pathways from identities of neighbor Pfam families

EFI-EST and EFI-GNT Web Tools: EFI-GNT



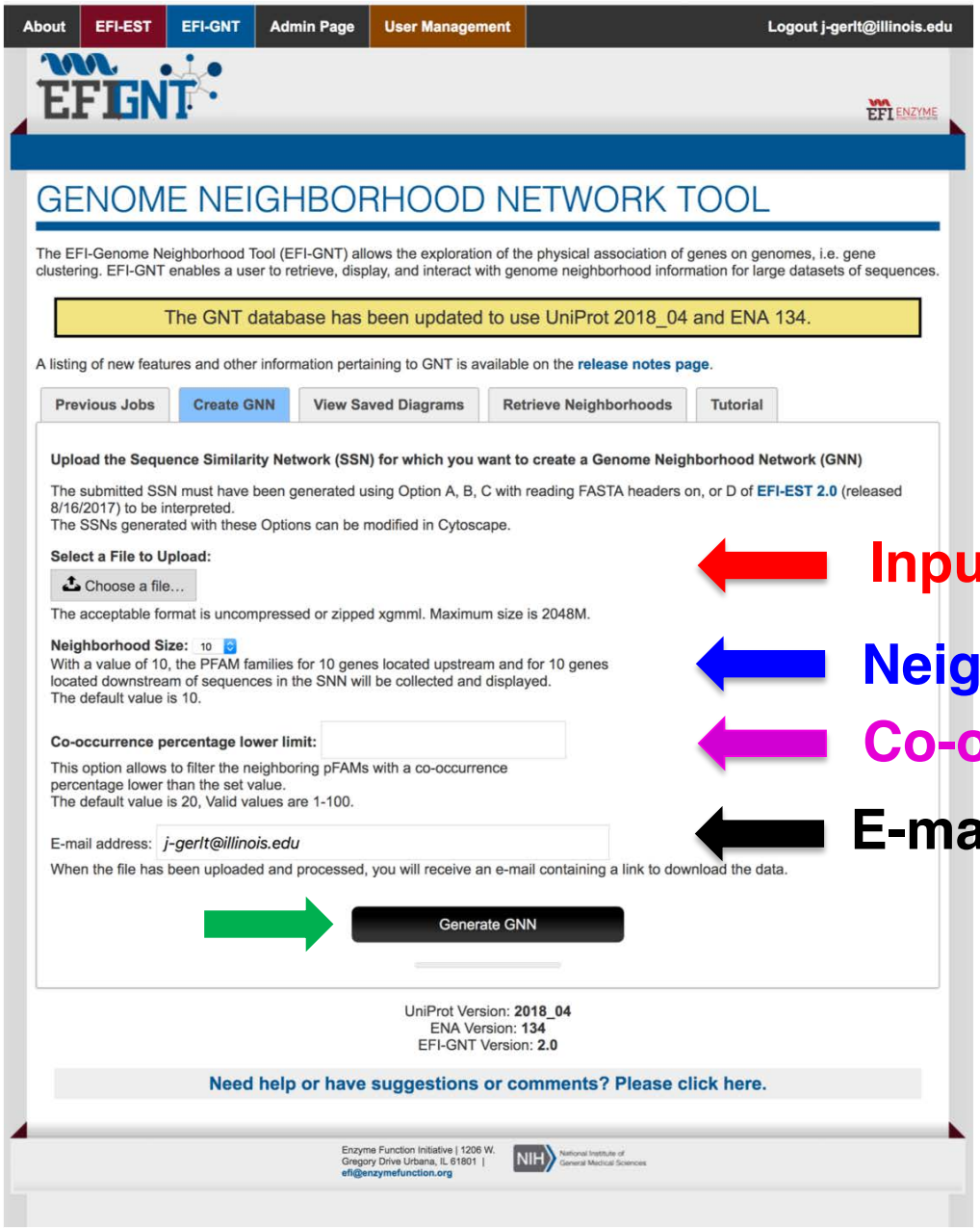
https://efi.igb.illinois.edu/efi-gnt/

EFI-GNT uses genome sequences from ENA (EMBL/EBI)

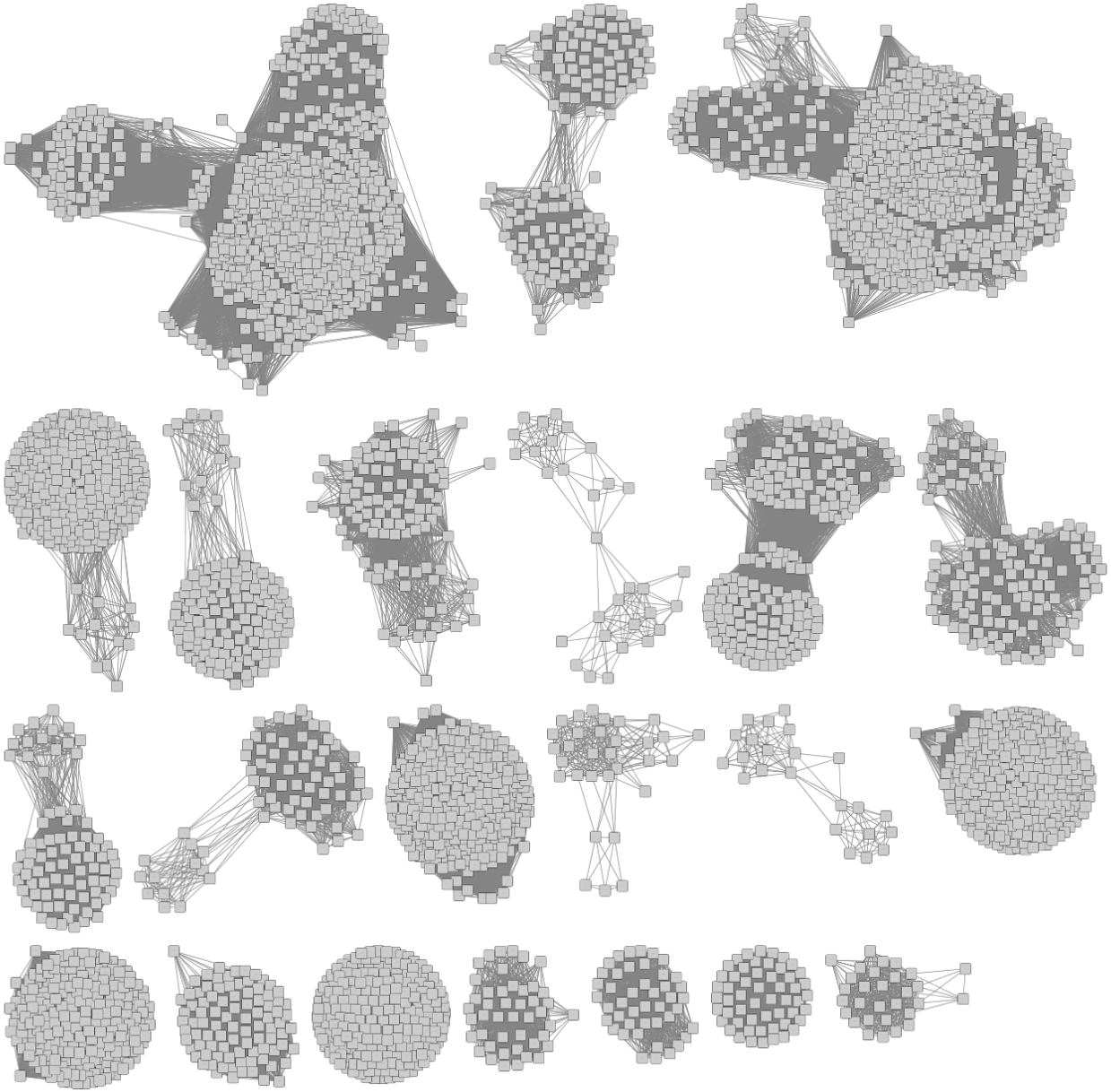


https://www.ebi.ac.uk/ena

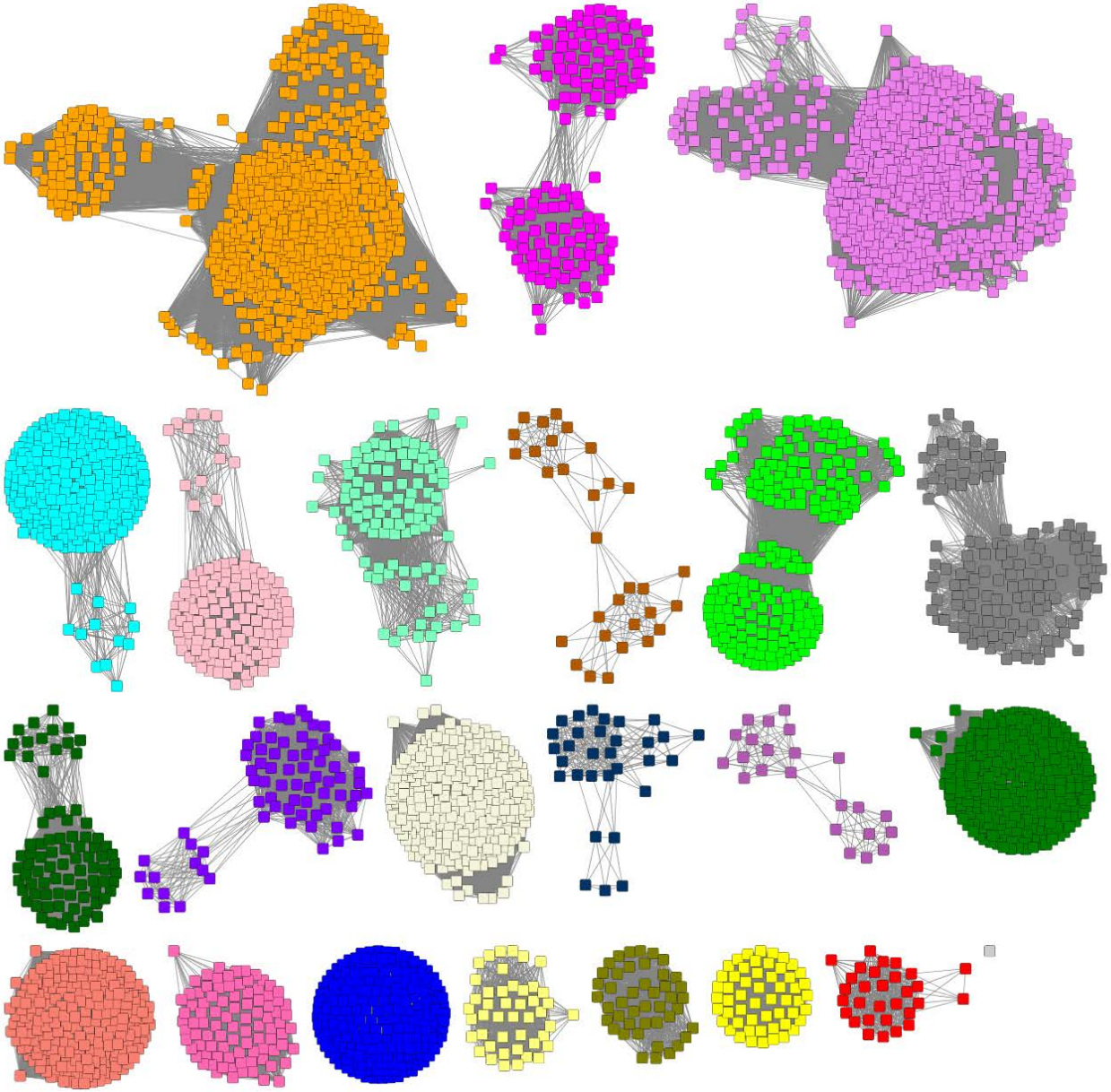
EFI-GNT: create GNN from SSN



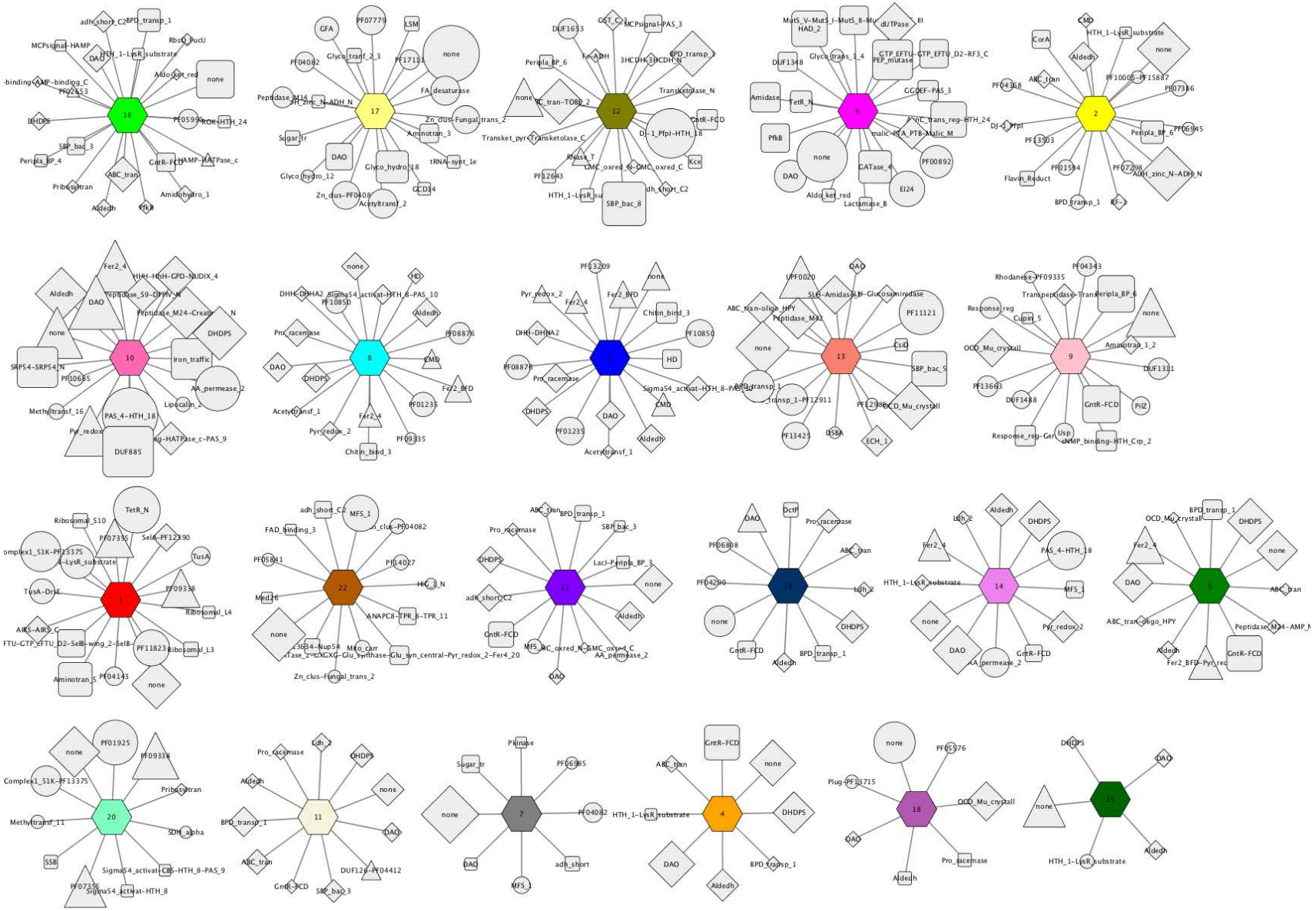
Input SSN: from EFI-EST



Colored SSN: families (queries) with unique colors

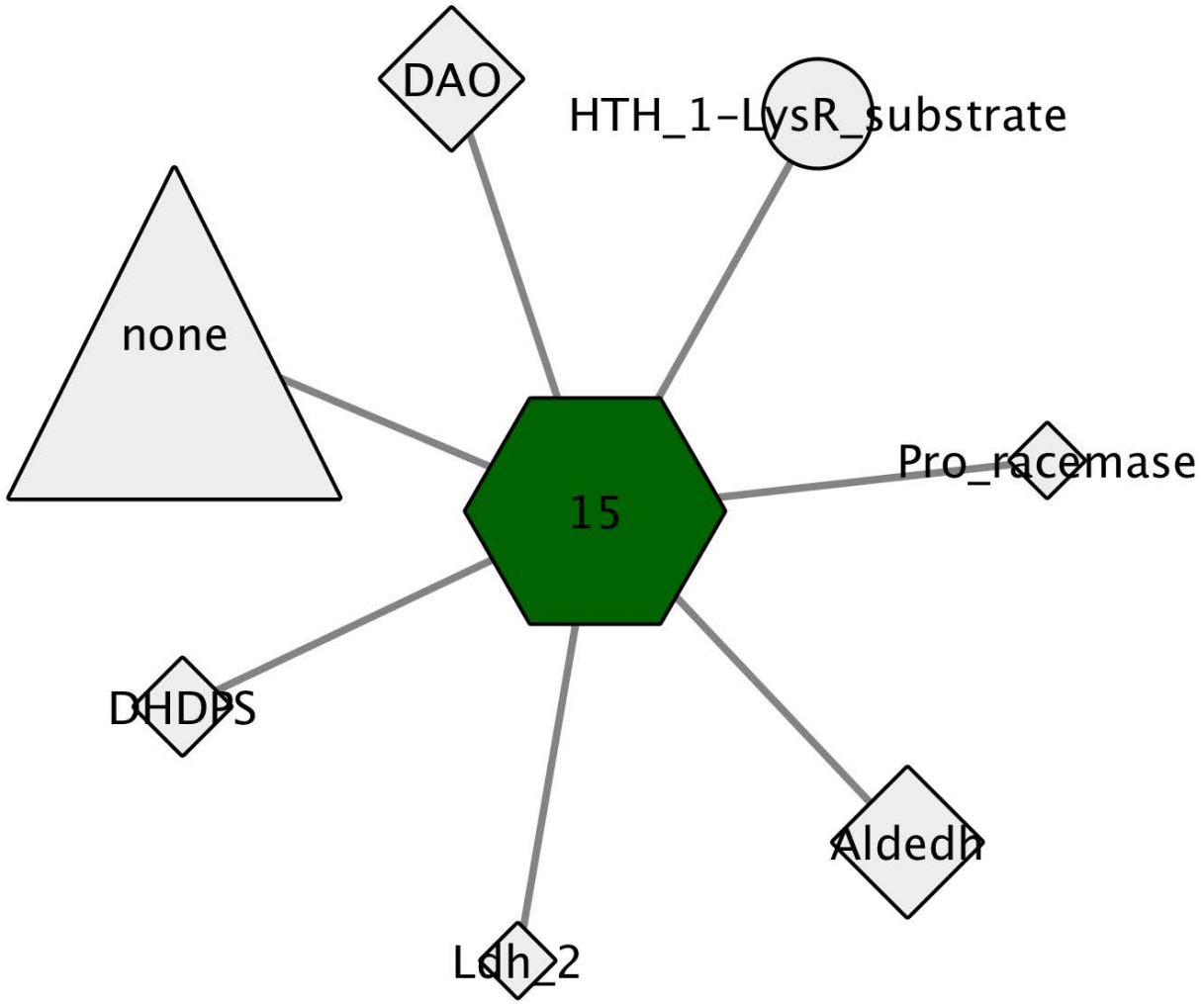


First GNN: Pfam families found by SSN clusters



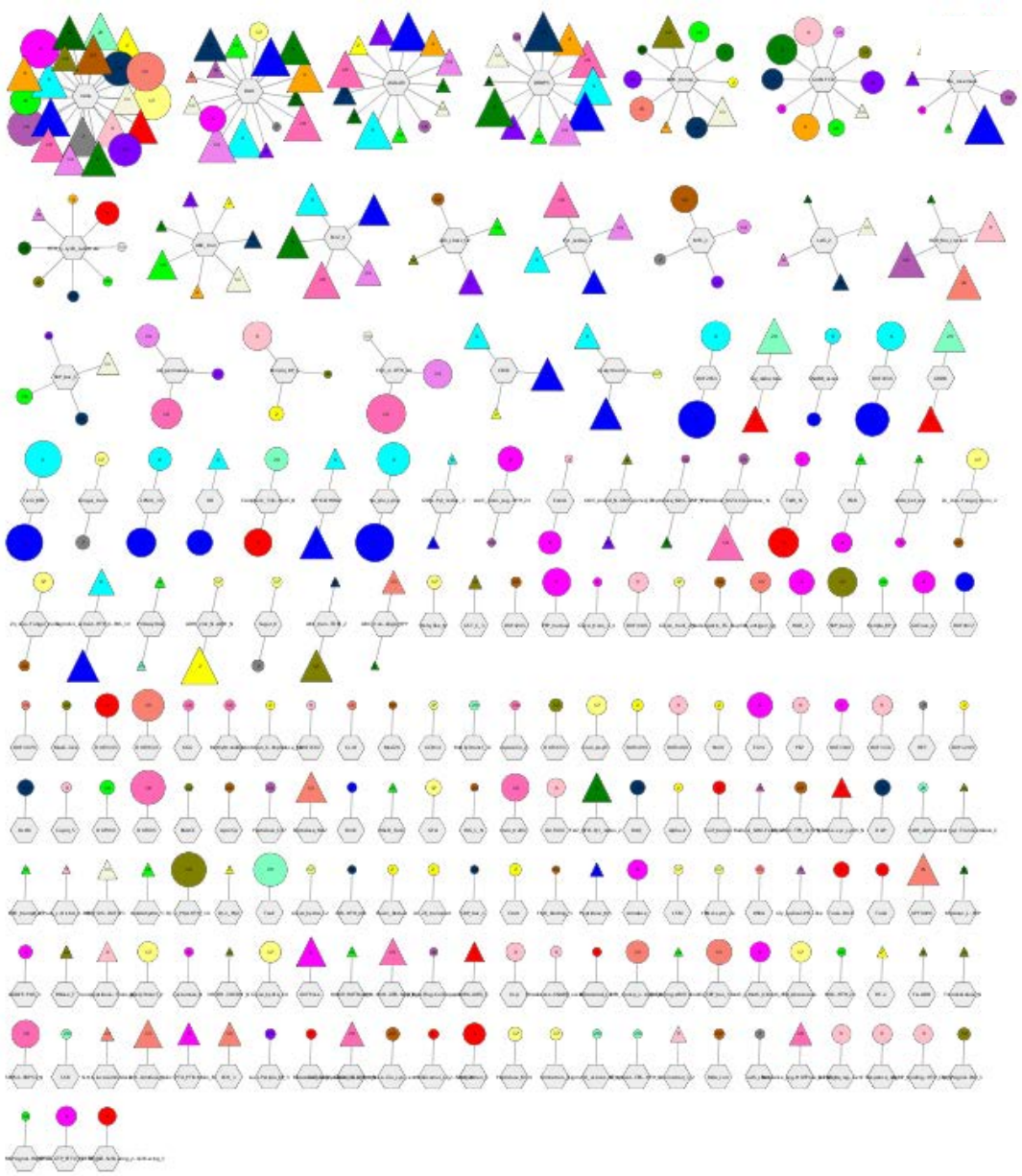
SSN cluster hub-nodes and Pfam family spoke-nodes

Functionally linked enzymes in pathways



Second GNN: SSN clusters found by Pfam families

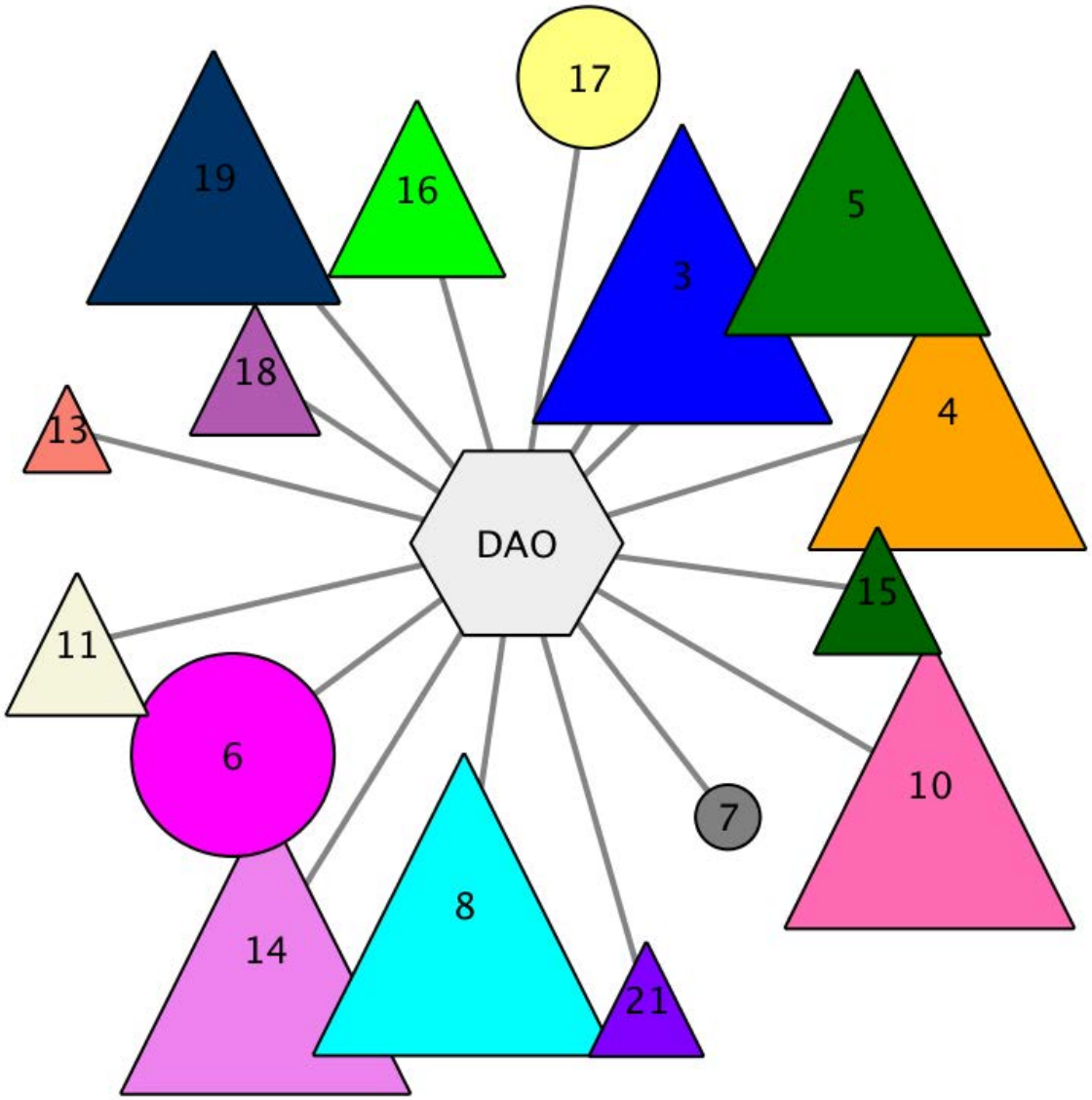
Shared context



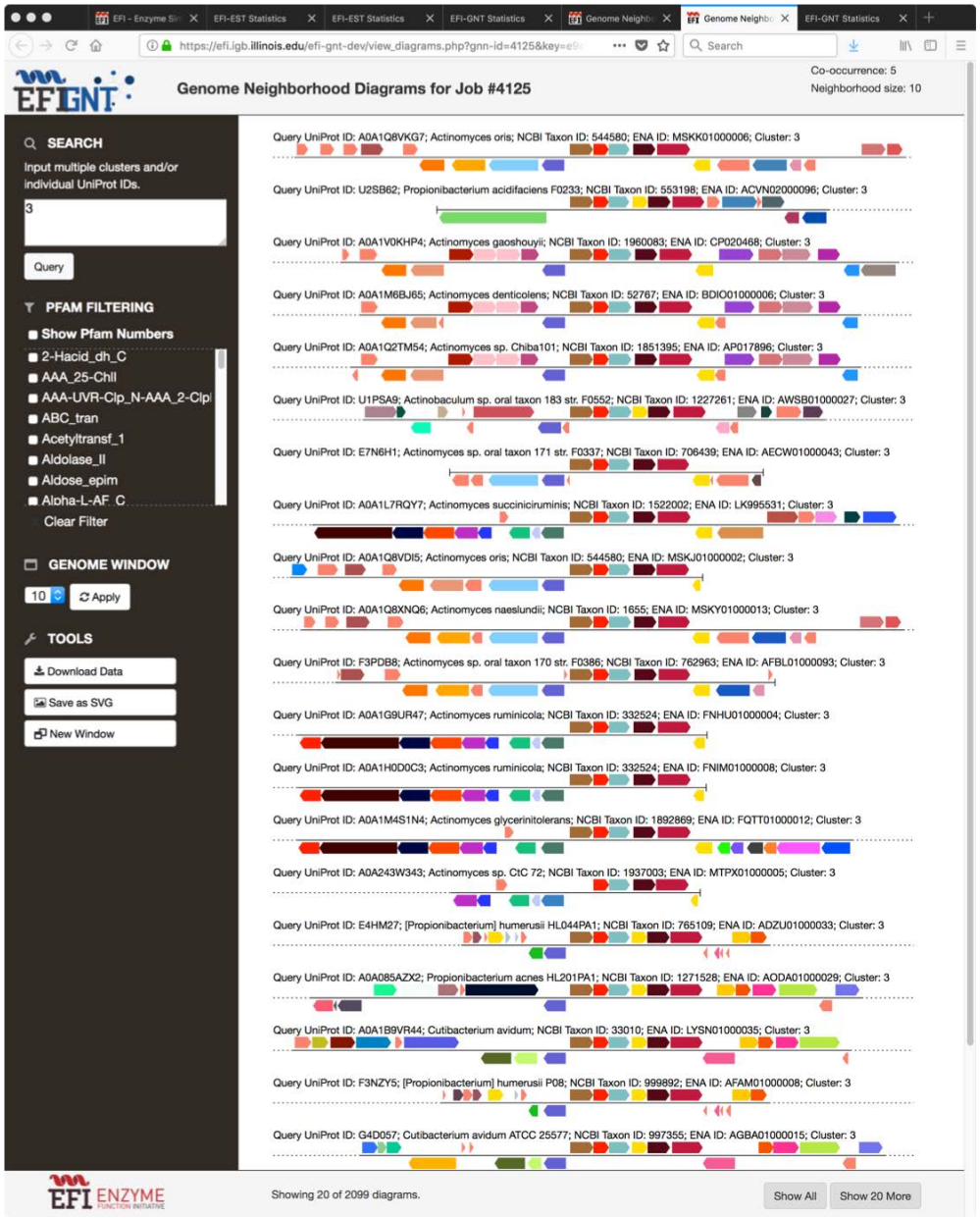
Unique context

Pfam family hub-nodes and SSN cluster spoke-nodes

SSN clusters with shared functions



Genome Neighborhood Diagrams



Today

1. Target selection using sequence similarity networks: EFI-EST web tool

2. Pathway discovery using genome neighborhood networks: EFI-GNT web tool

3. **Example: using SSNs and GNNs to discover catabolic pathways for D-apiose**

4. Target prioritization using chemically guided functional profiling: CGFP-ShortBRED beta tool

Functional assignment of multiple catabolic pathways for D-apiose

***Nature Chemical Biology*, in press**
DOI: 10.1038/s41589-018-0067-7
Available on line, June 4

A general strategy for discovering pathways

BIOCHEMISTRY
including biophysical chemistry & molecular biology

Article

pubs.acs.org/biochemistry

Experimental Strategies for Functional Annotation and Metabolism Discovery: Targeted Screening of Solute Binding Proteins and Unbiased Panning of Metabolomes

Matthew W. Vetting,[*,†,¶] Nawar Al-Obaidi,[†,¶] Suwen Zhao,[‡,¶] Brian San Francisco,[§,¶] Jungwook Kim,[†] Daniel J. Wichelecki,[§,∥] Jason T. Bouvier,[§,∥] Jose O. Solbiati,[§] Hoan Vu,[#] Xinshuai Zhang,[§] Dmitry A. Rodionov,[○,▽] James D. Love,[†] Brandan S. Hillerich,[†] Ronald D. Seidel,[†] Ronald J. Quinn,[*,#] Andrei L. Osterman,[*,○] John E. Cronan,[*,∥,◆] Matthew P. Jacobson,[*,‡] John A. Gerlt,[*,§,∥,⊥] and Steven C. Almo[*,†]

[†]Department of Biochemistry, Albert Einstein College of Medicine, Bronx, New York 10461, United States

[‡]Department of Pharmaceutical Chemistry, University of California, San Francisco, California 94158, United States

[§]Institute for Genomic Biology, [∥]Department of Biochemistry, [⊥]Department of Chemistry, [◆]Department of Microbiology, University of Illinois at Urbana−Champaign, Urbana, Illinois 61801, United States

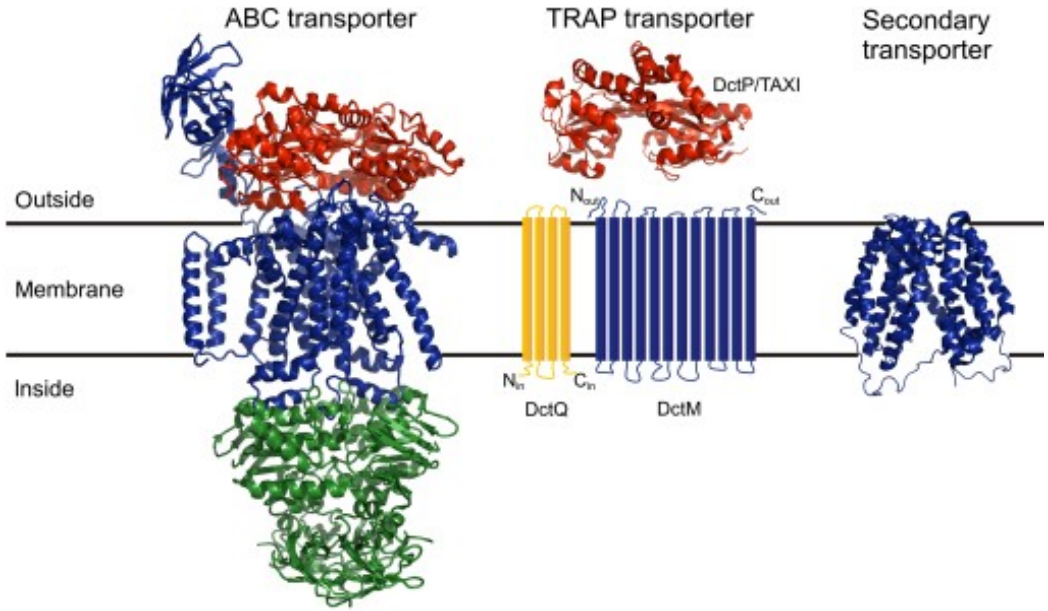
[#]Eskitis Institute for Drug Discovery, Griffith University, Brisbane, Queensland 4111, Australia

[○]Sanford-Burnham Medical Research Institute, La Jolla, California 92037, United States

[▽]A.A.Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow 127994, Russia

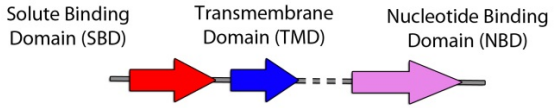
DOI: 10.1021/bi501388y
Biochemistry 2015, 54, 909−931

Transporter Solute Binding Proteins (SBPs): genes colocated with those for catabolic pathways

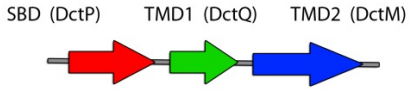


3B50
TRAP SBP
Sialic acid
Binding Protein

ABC TRANSPORTERS
(ATP BINDING CASSETTE)



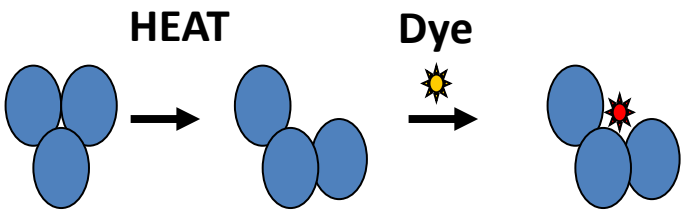
TRAP TRANSPORTERS
(**TR**ipartite **A**TP-independent **P**eriplasmic transporters)

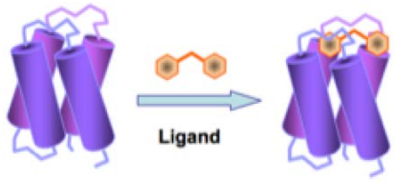


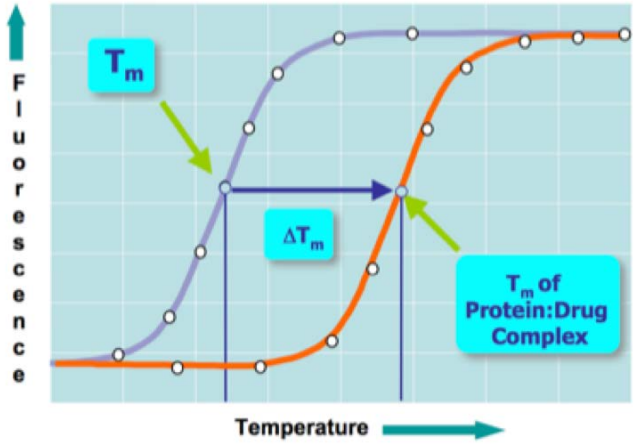
1ANF
ABC SBP
Maltose
Binding Protein

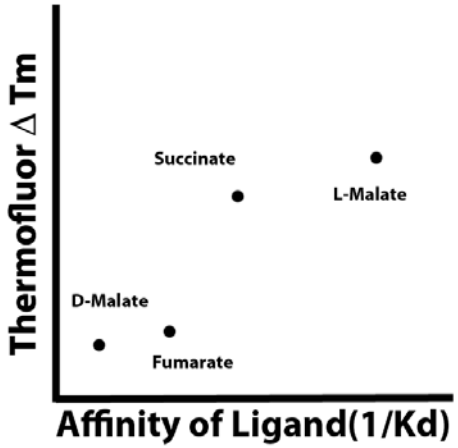
Mulligan, C et al. (2011). "Tripartite ATP-independent periplasmic (TRAP) transporters in bacteria and archaea." FEMS Microbiol Rev **35**(1): 68-86.

Ligand screening by differential scanning fluorimetry (DSF or ThermoFluor)









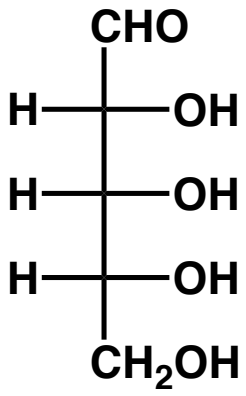
Order of ligand stabilization correlated with affinity of ligand but not absolute

Applied Biosystems 7900HT
Fast Real-Time PCR System

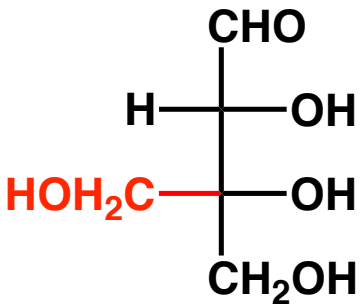
1 mM Ligand
10 µm Protein (100X)
100 mM Hepes pH 7.5, 150 mM NaCl
Sypro Orange (Indicator Dye)

Library Hits for ABC Transport SBPs (PF13407)

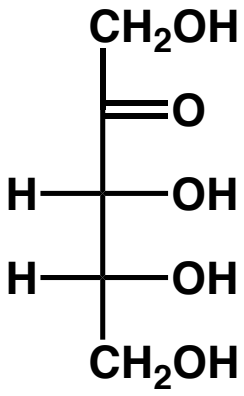
SBP (UniProt)	Organism	D-Ribose (ΔTm °C)	D-Apiose (ΔTm °C)	D-Ribulose (ΔTm °C)
Q2JZQ5	*Rhizobium etli* CFN 42	12.5	**6.8**	3.3
B1G898	*Burkholderia graminis* C4D1M	15.9	**8.8**	4.7
A6VKQ8	*Actinobacillus succinogenes* 130Z	16.2	**9.7**	1.4



D-Ribose

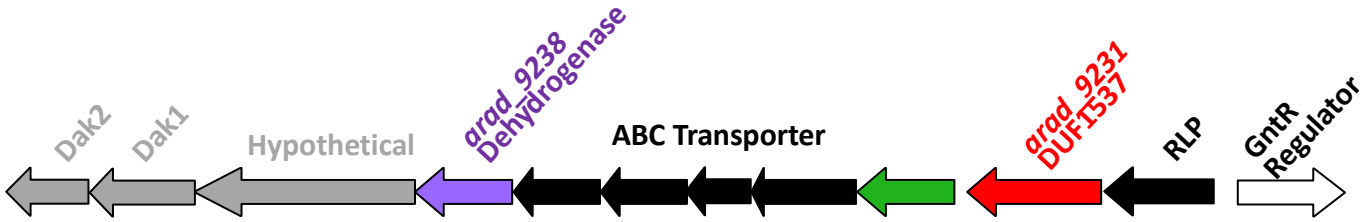


D-Apiose

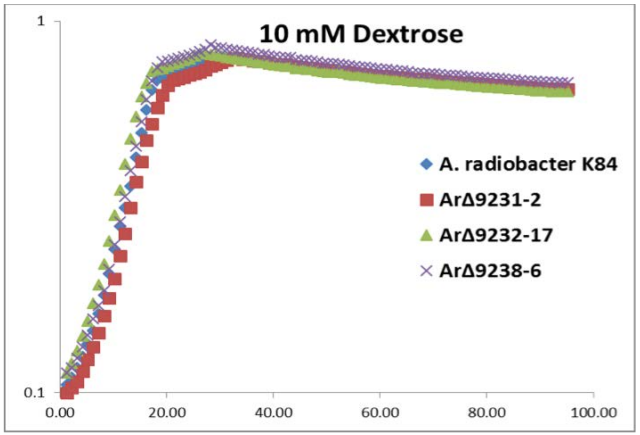


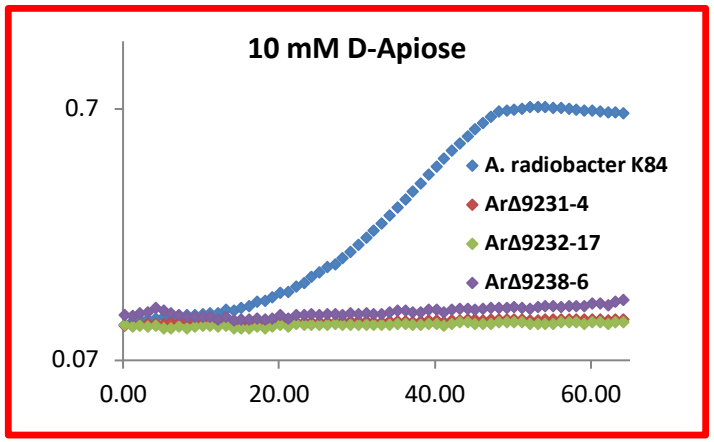
D-Ribulose

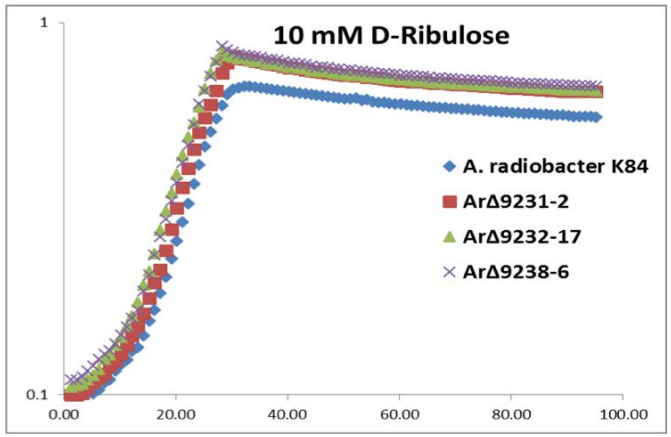
Knockouts in *Agrobacterium radiobacter*: D-apiose

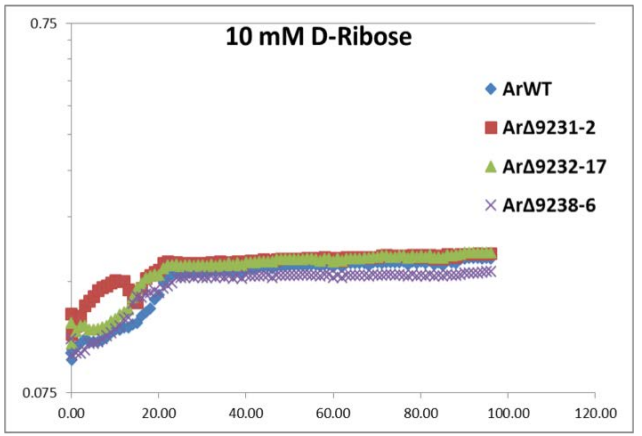


***Agrobacterium radiobacter* K84**



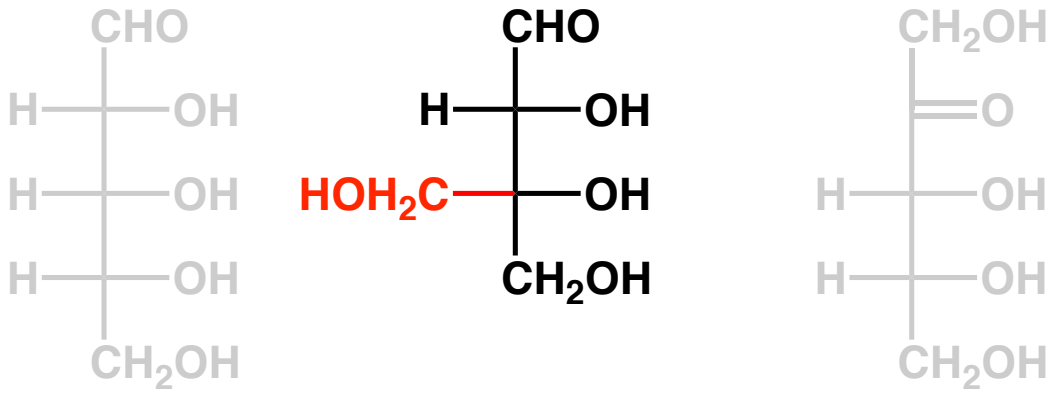




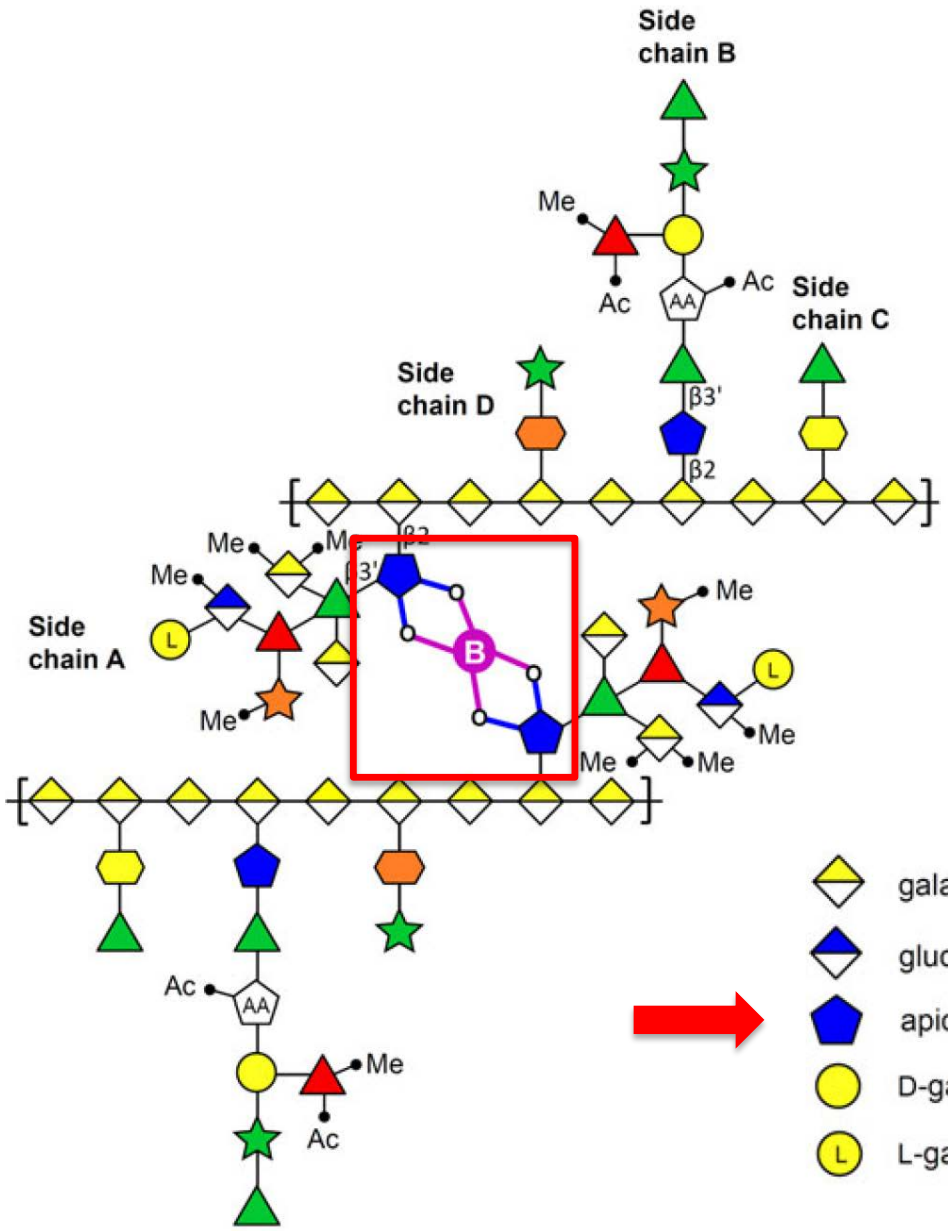


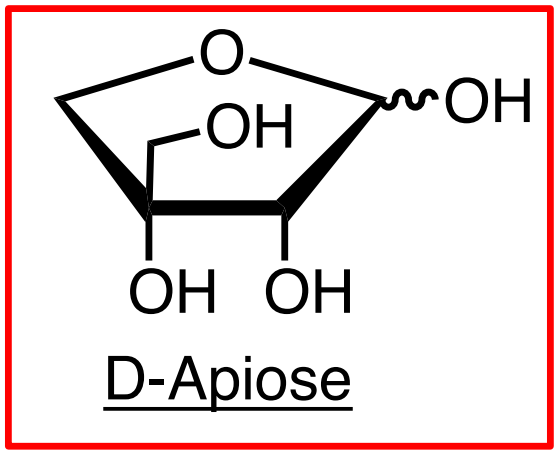
D-Apiose is physiological ligand

SBP (UniProt)	Organism	D-Ribose (ΔTm °C)	D-Apiose (ΔTm °C)	D-Ribulose (ΔTm °C)
Q2JZQ5	*Rhizobium etli* CFN 42	12.5	**6.8**	3.3
B1G898	*Burkholderia graminis* C4D1M	15.9	**8.8**	4.7
A6VKQ8	*Actinobacillus succinogenes* 130Z	16.2	**9.7**	1.4



Rhamnogalacturonan-II: D-apiose cross-links in cell walls





galacturonic acid
glucuronic acid
apiose
D-galactose
L-galactose
rhamnose
fucose
arabinose
xylose
aceric acid
DHA
KDO
Me methyl
Ac acetyl
B borate

Glycoside hydrolases and lyases: release D-apiose

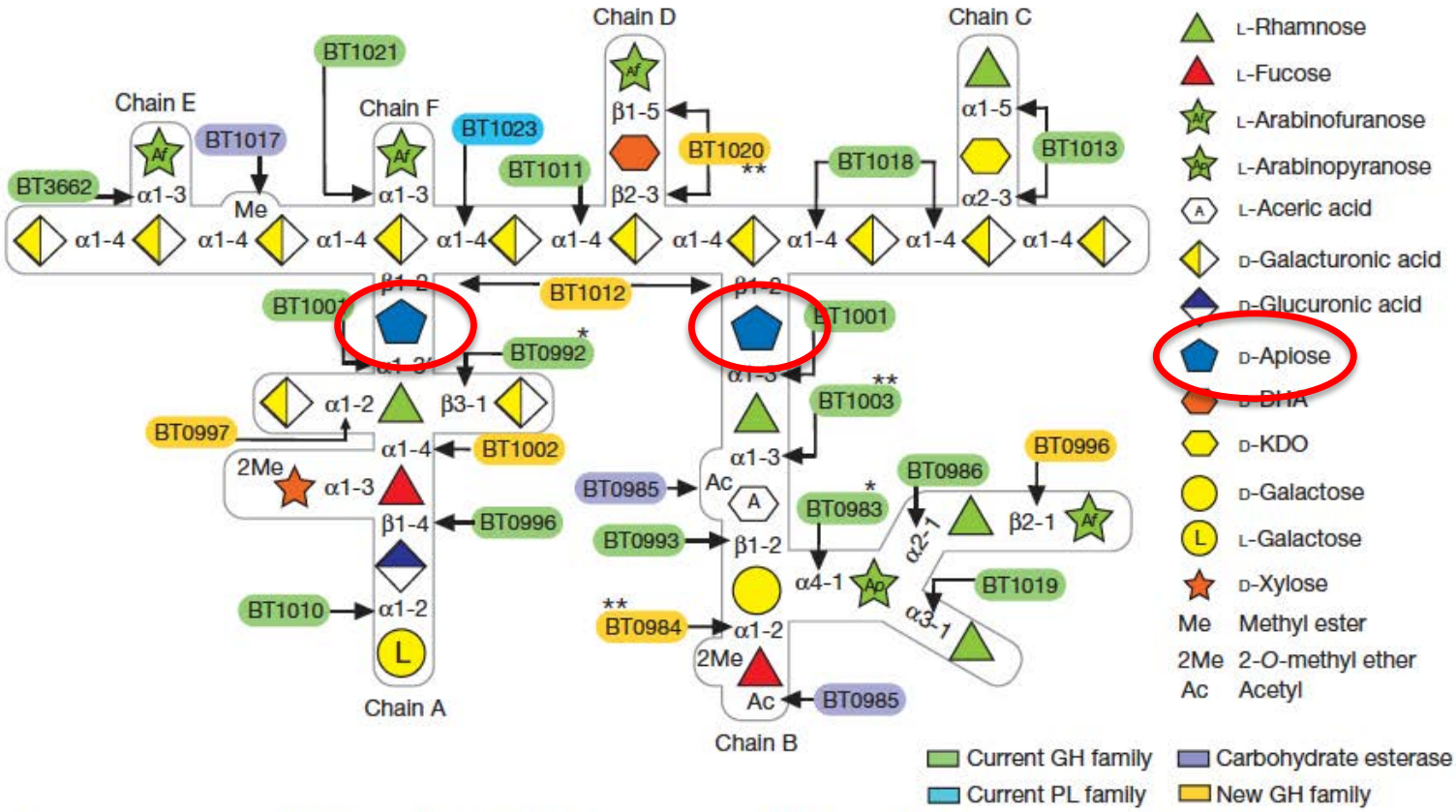


Figure 1 | Schematic of enzymes and PULs involved in RG-II degradation. Sugars shown using the Consortium for Functional Glycomics notation[25]. Enzymes are appropriately colour-coded. Single asterisks denote a new activity for a glycoside hydrolase (GH) family; double asterisks denote enzymes with activities that have not previously been observed. D-KDO, 3-deoxy-D-manno-octulosonic acid; D-DHA, 2-keto-3-deoxy-d-*lyxo*-heptulosaric acid; PL, polysaccharide lyase.

But, no catabolic pathway for D-apiose ?

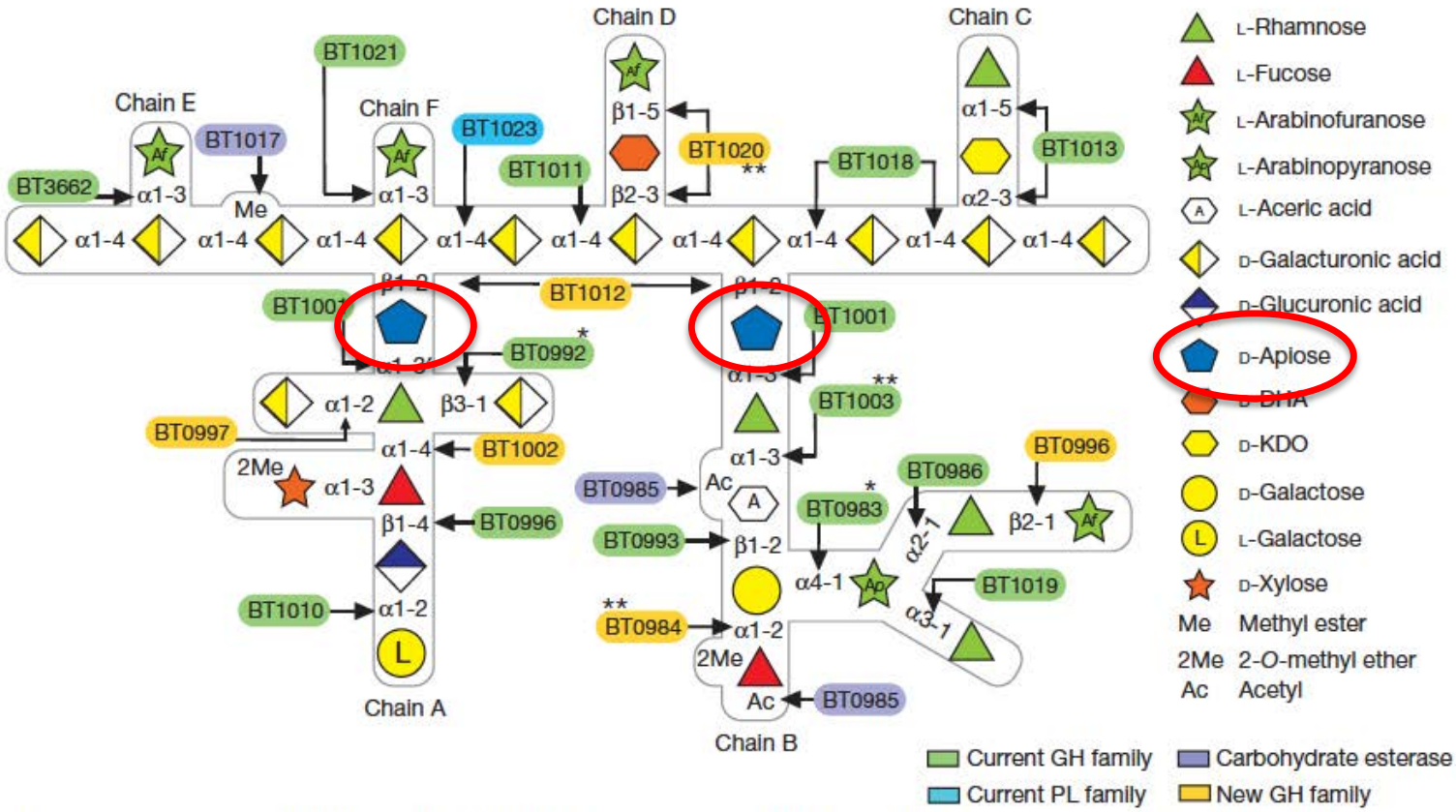
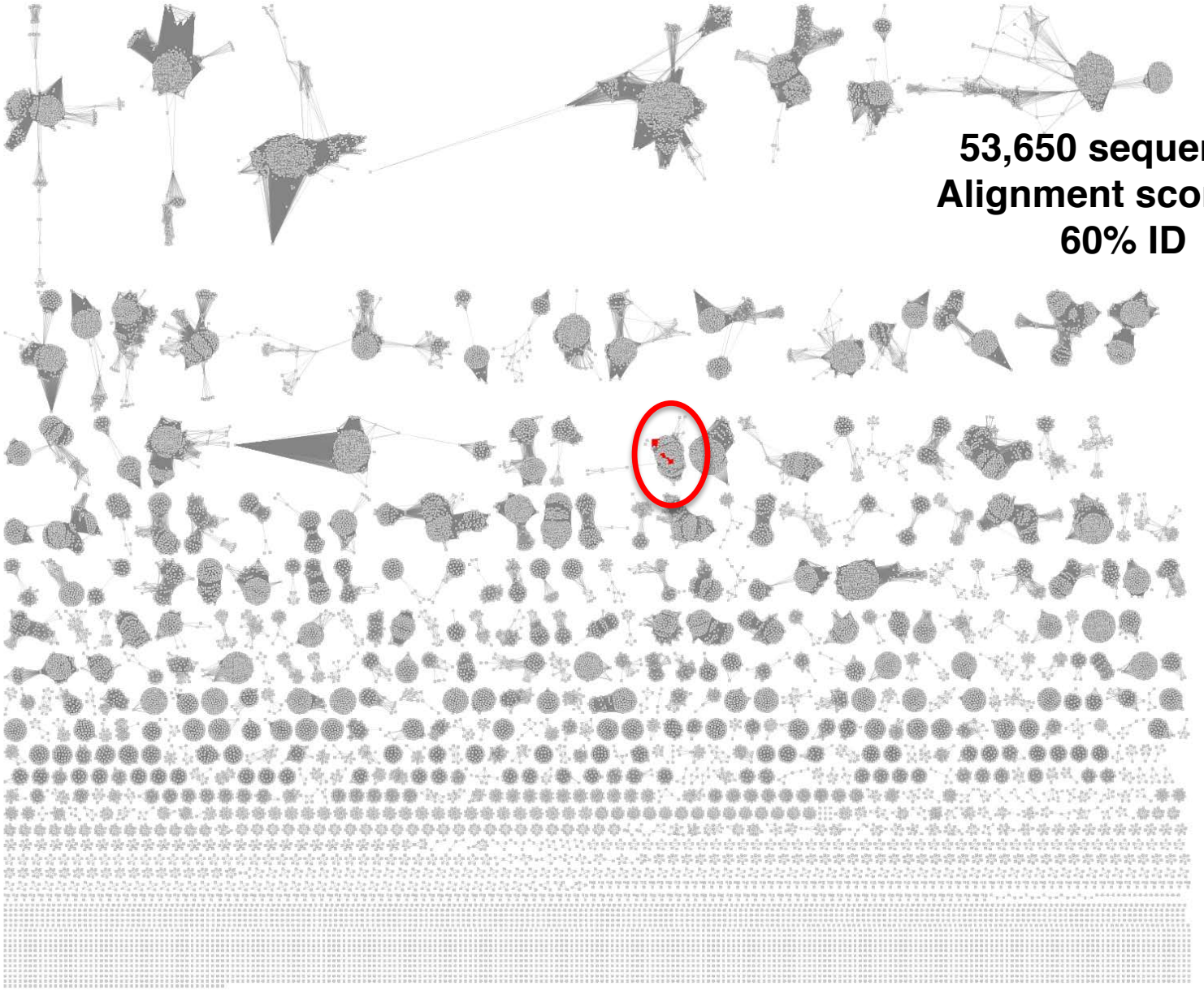


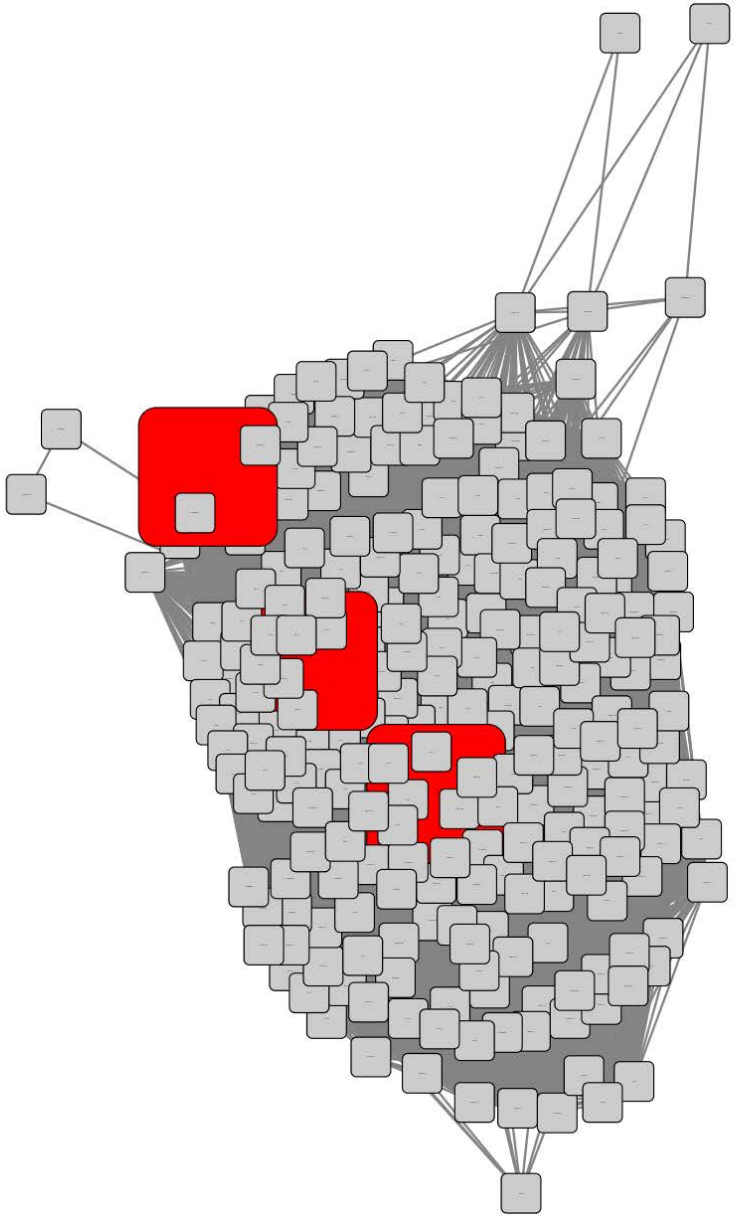
Figure 1 | Schematic of enzymes and PULs involved in RG-II degradation. Sugars shown using the Consortium for Functional Glycomics notation[25]. Enzymes are appropriately colour-coded. Single asterisks denote a new activity for a glycoside hydrolase (GH) family; double asterisks denote enzymes with activities that have not previously been observed. D-KDO, 3-deoxy-D-manno-octulosonic acid; D-DHA, 2-keto-3-deoxy-d-*lyxo*-heptulosaric acid; PL, polysaccharide lyase.

D-Apiose: three hits in one SBP cluster



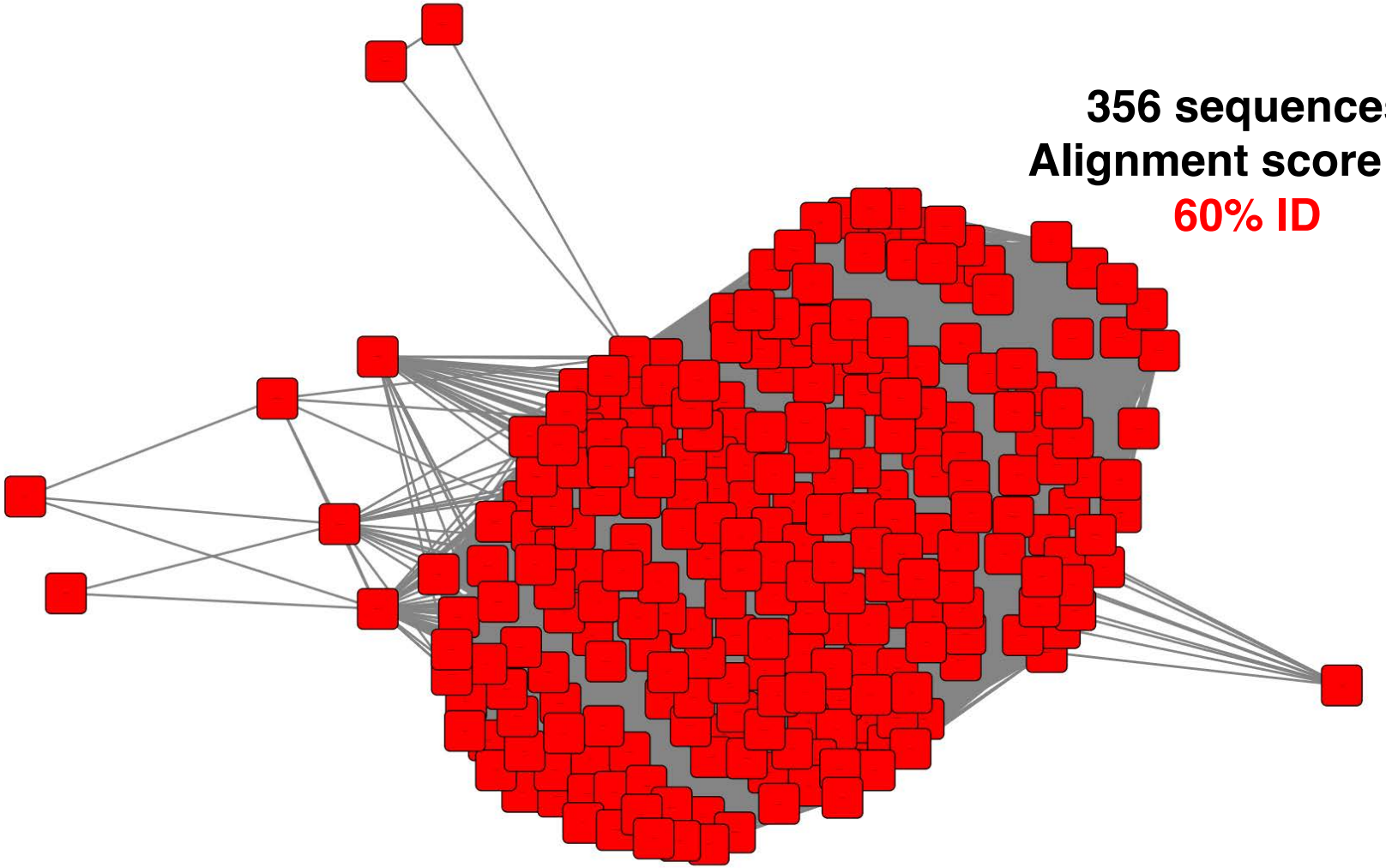
53,650 sequences
Alignment score 110
60% ID

D-apiose SBP cluster

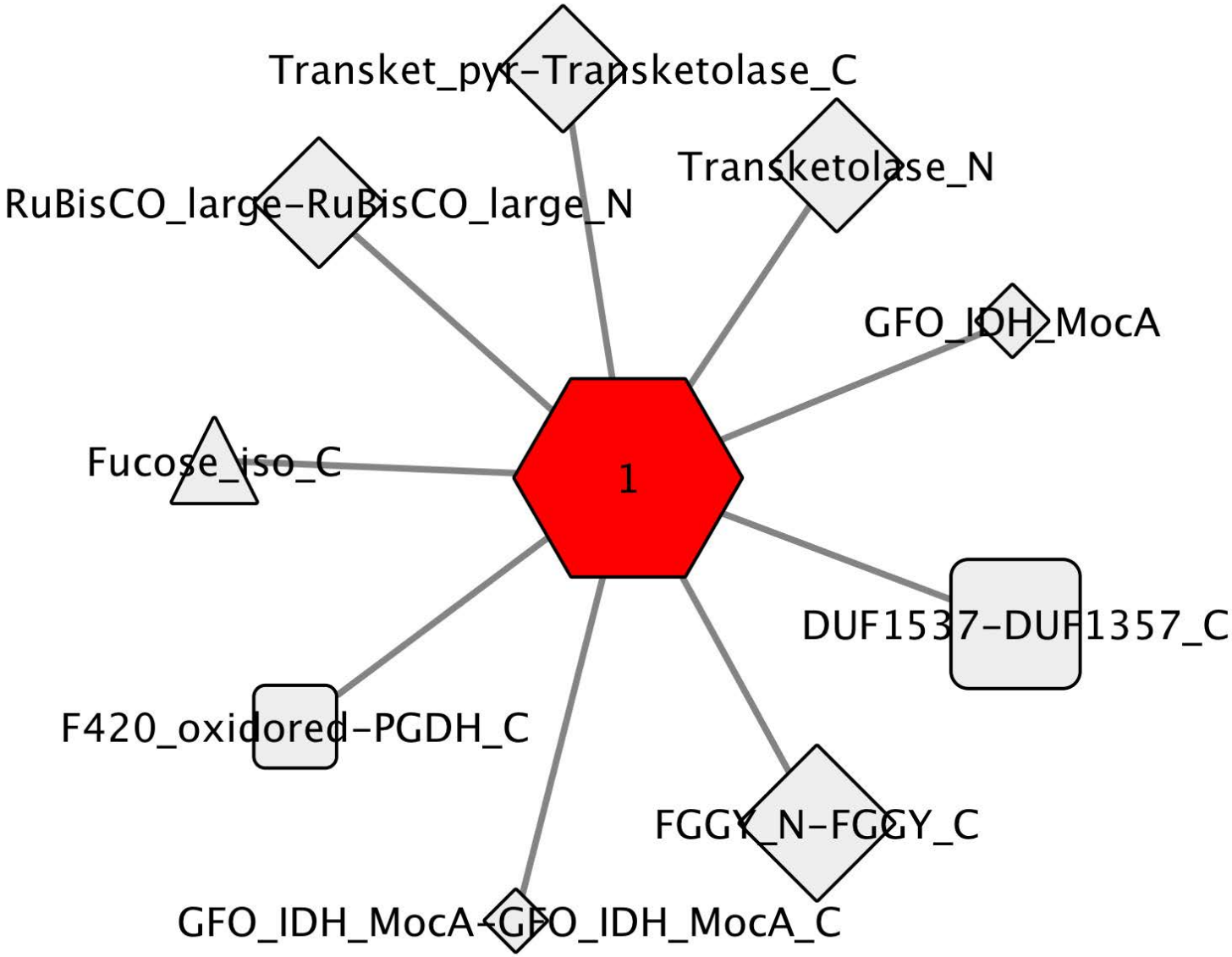


356 sequences
Alignment score 110
60% ID

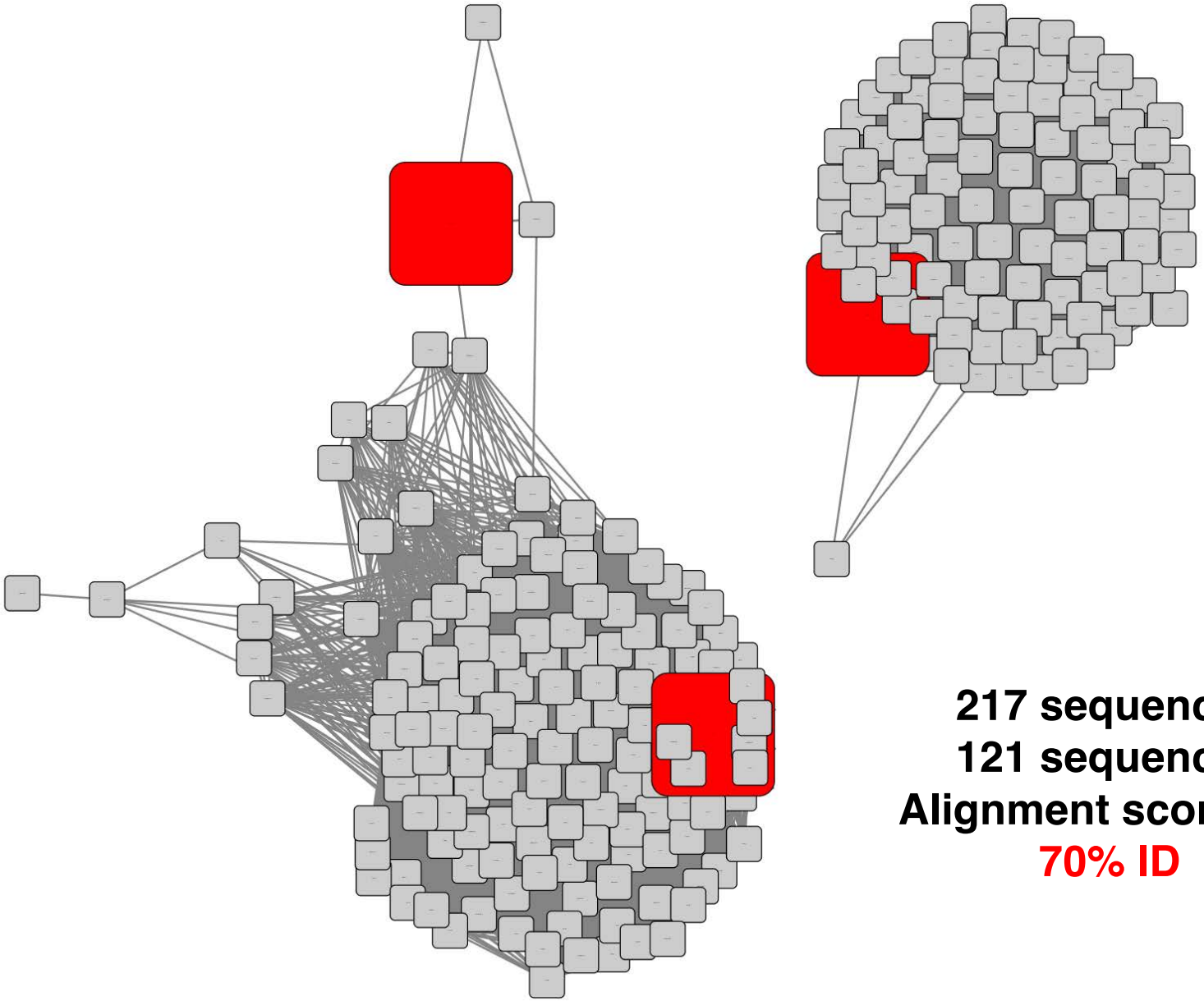
D-Apiose SBP hits: colored SSN



GNN: (too) many enzyme families, multiple pathways ?

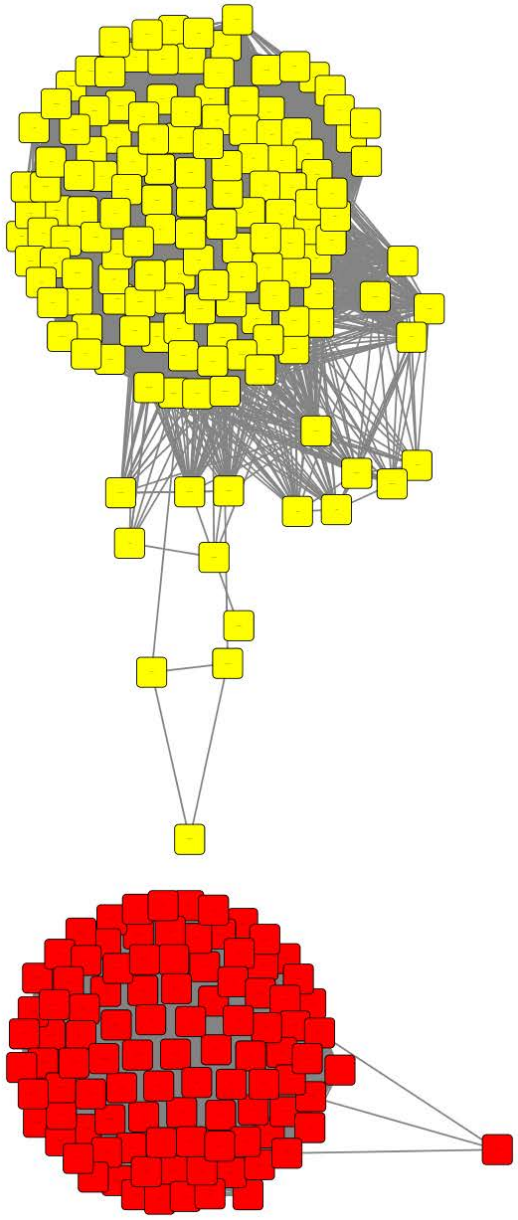


70% ID, alignment score 124: two clusters



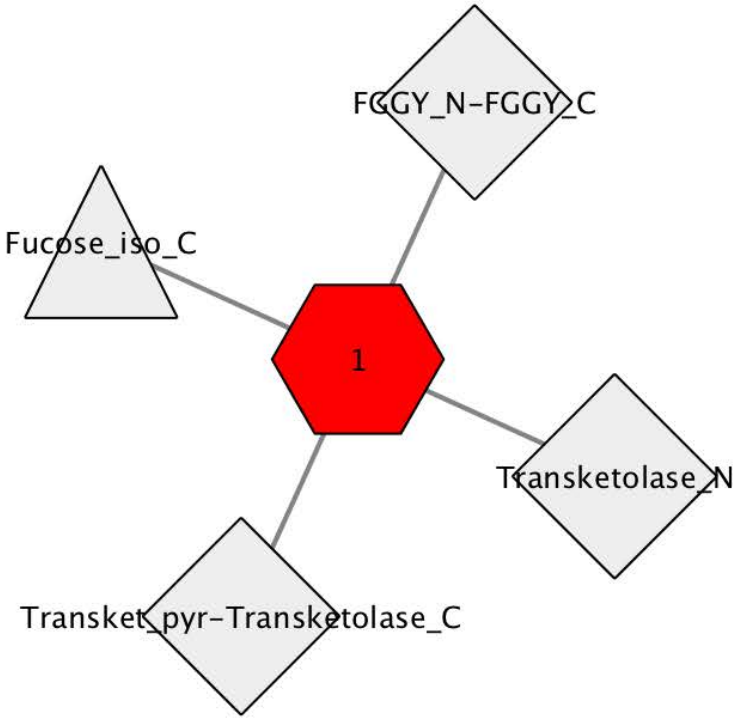
217 sequences
121 sequences
Alignment score 124
70% ID

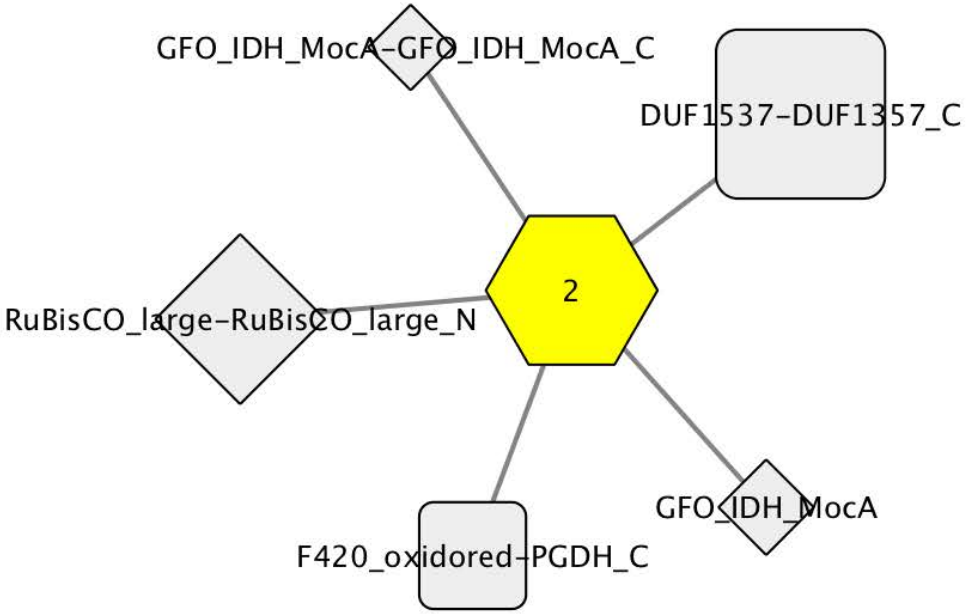
70% ID, alignment score 124: colored SSN



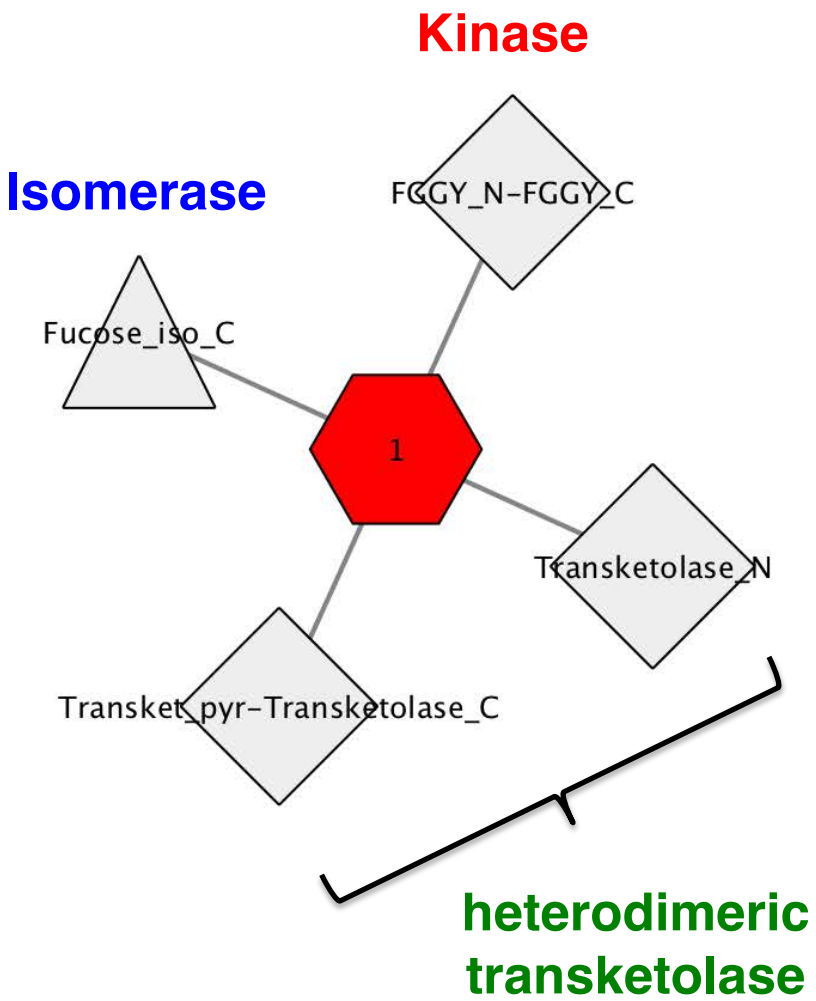
217 sequences
121 sequences
Alignment score 124
70% ID

GNN: two pathways ?

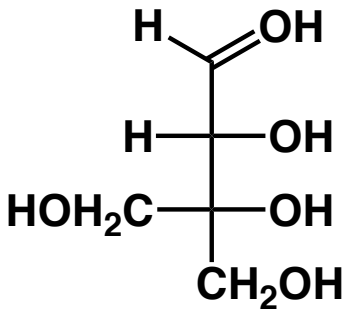




Red cluster: transketolase pathway

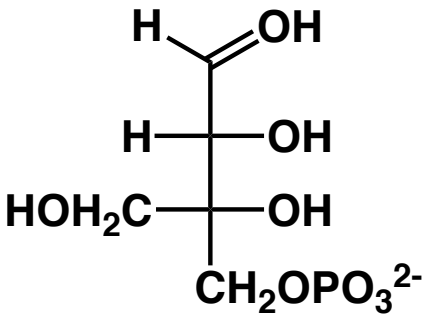


Transketolase pathway

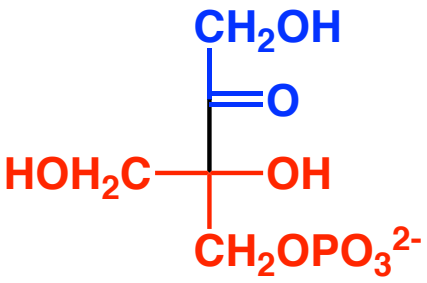


D-Apiose

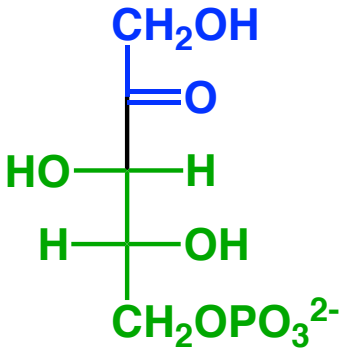
FGGY Kinase



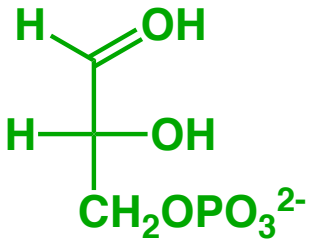
Isomerase

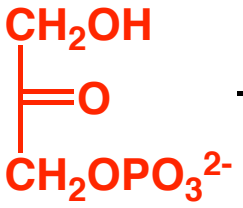


Transketolase



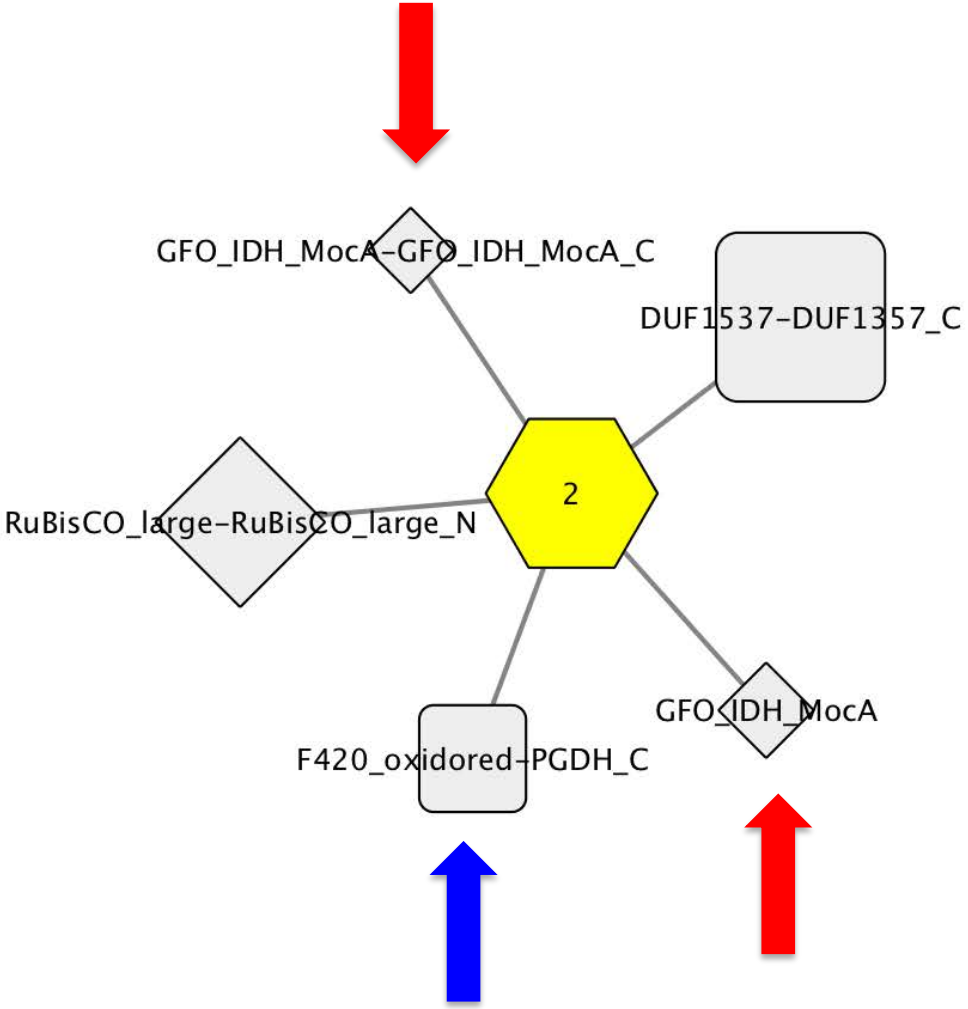
Pentose Phosphate Pathway



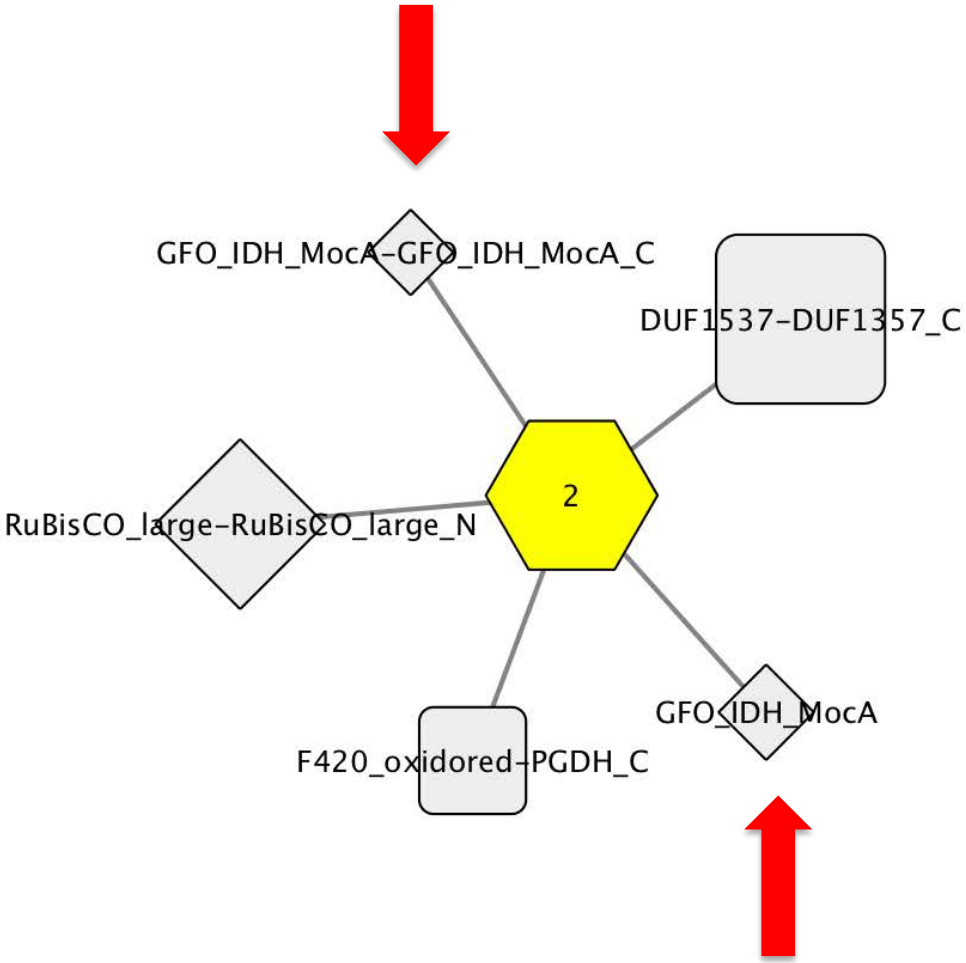


Glycolysis

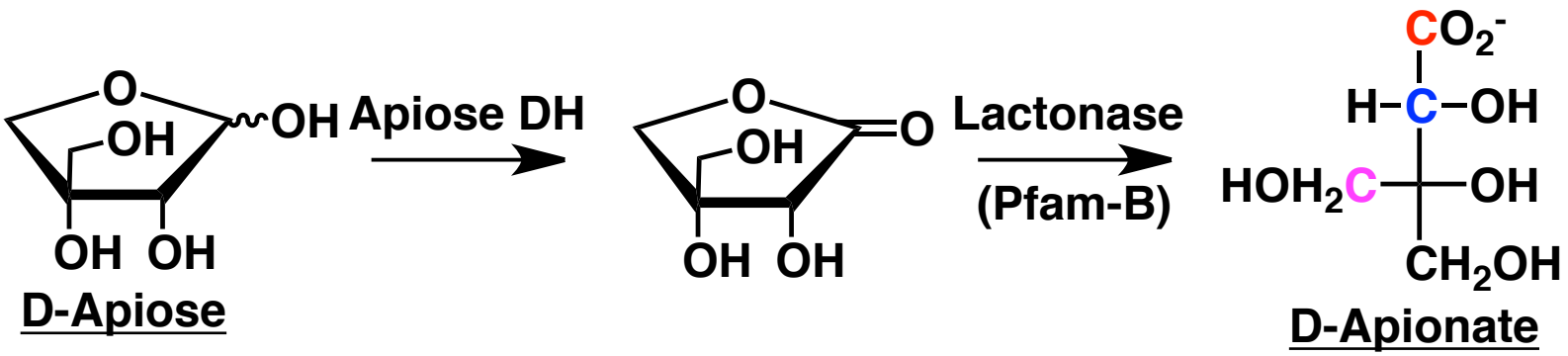
Yellow cluster: oxidative pathway(s) with two dehydrogenases



Oxidative pathway: D-apiose dehydrogenase



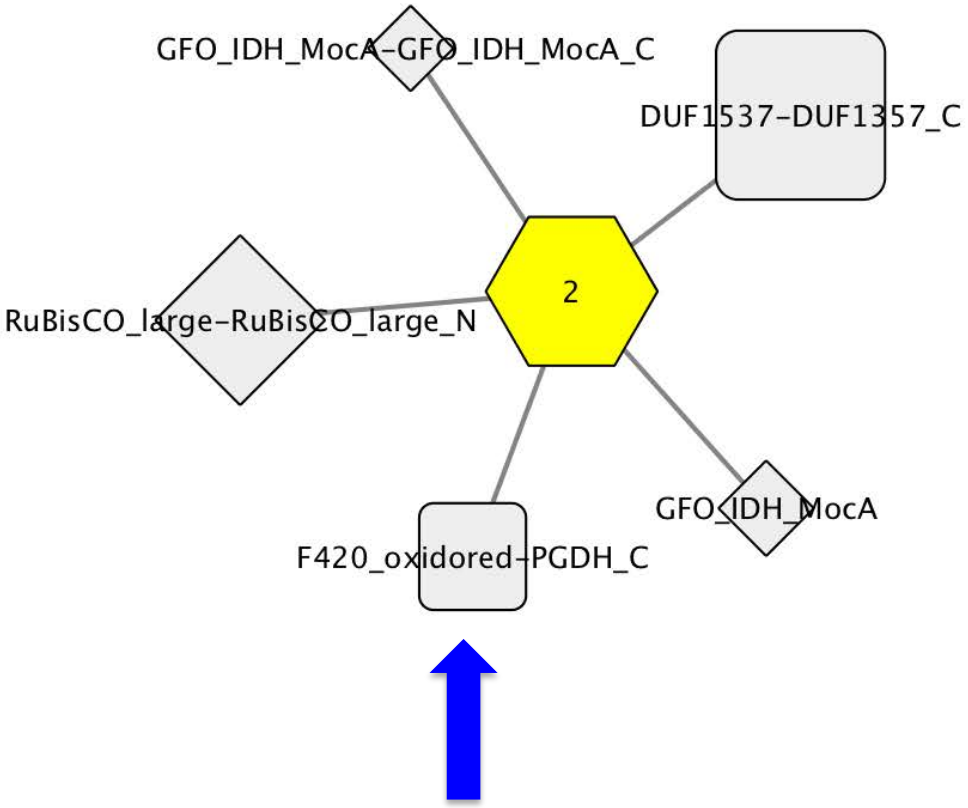
Oxidation of D-apionate



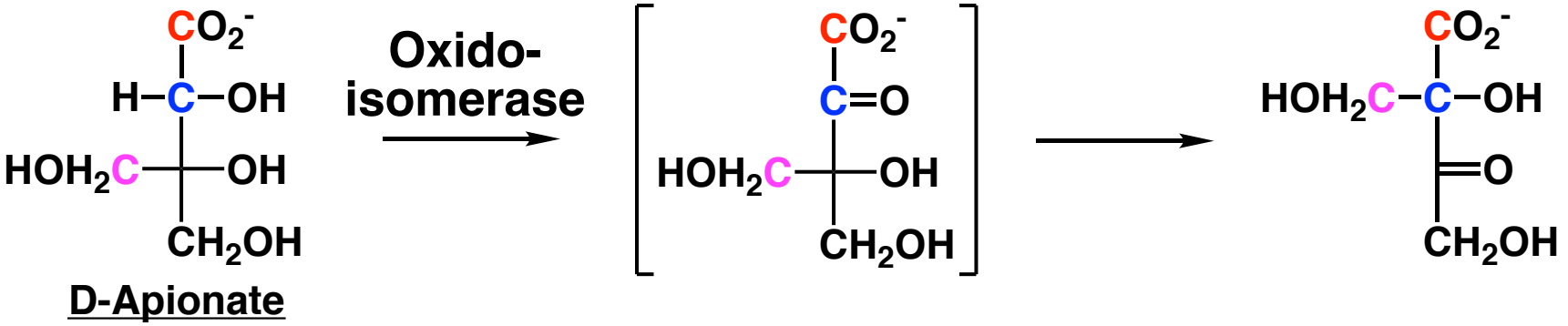
D-Apiose

D-Apionate

Oxidative pathway: oxidoisomerase

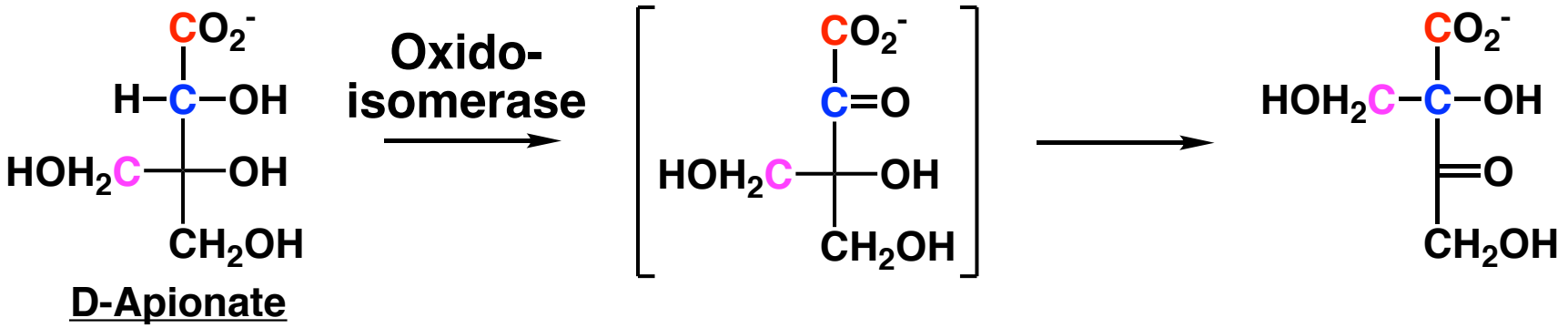


PF16896: novel oxidoisomerase

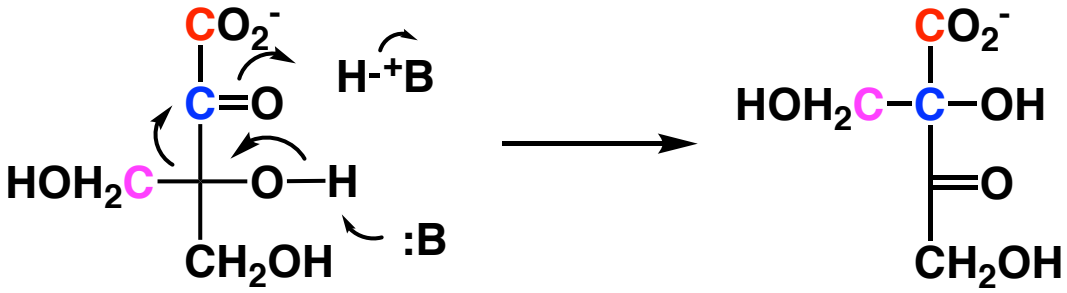


D-Apionate

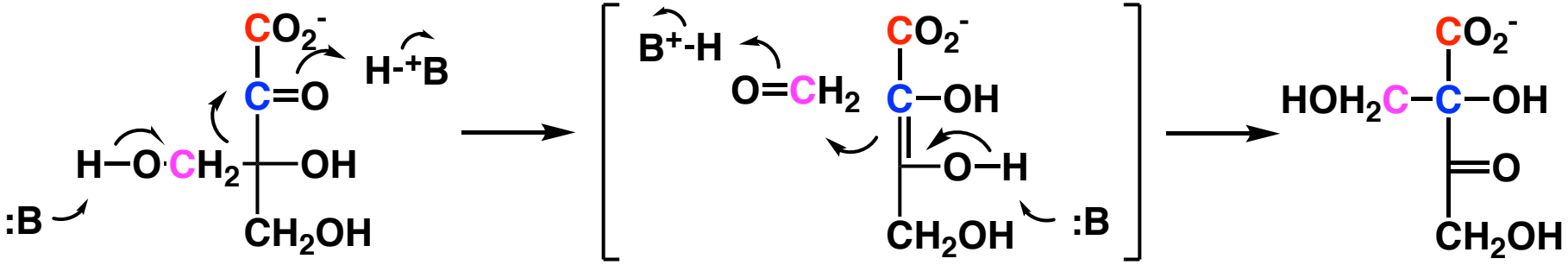
PF16896: novel oxidoisomerase



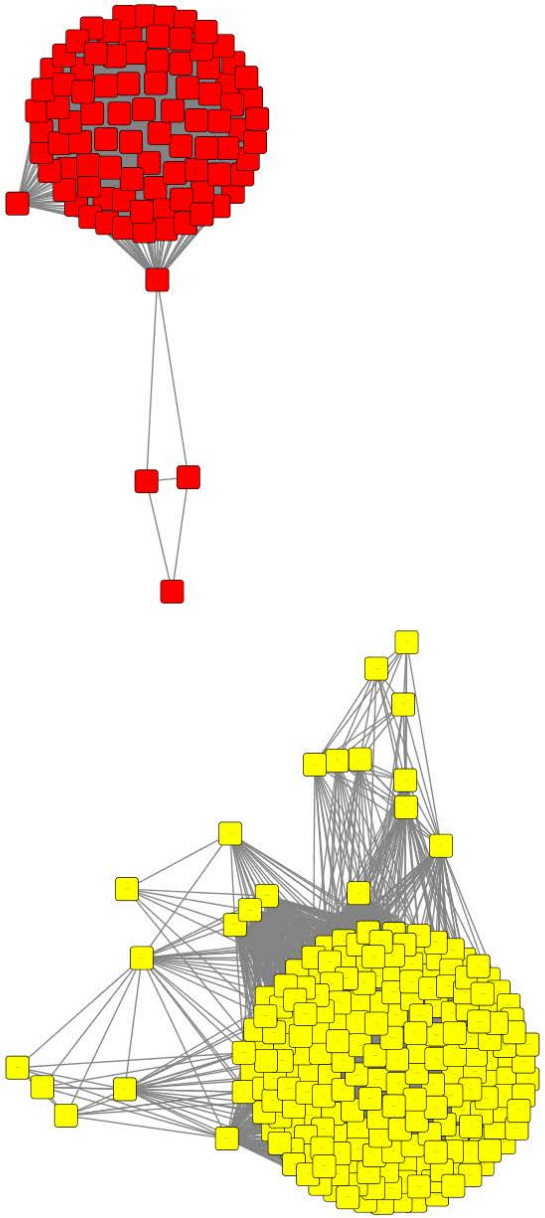
Concerted ?



Or, stepwise ?

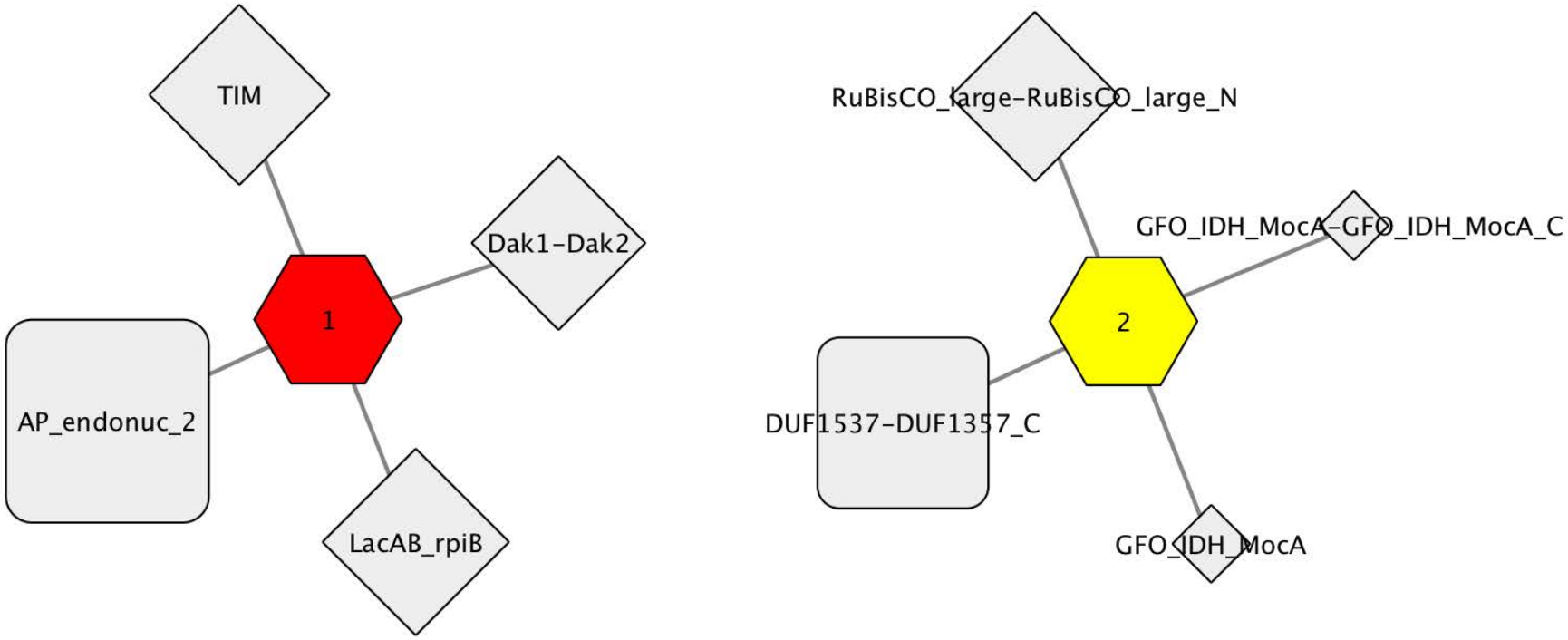


Oxidoisomerase SSN: two major clusters

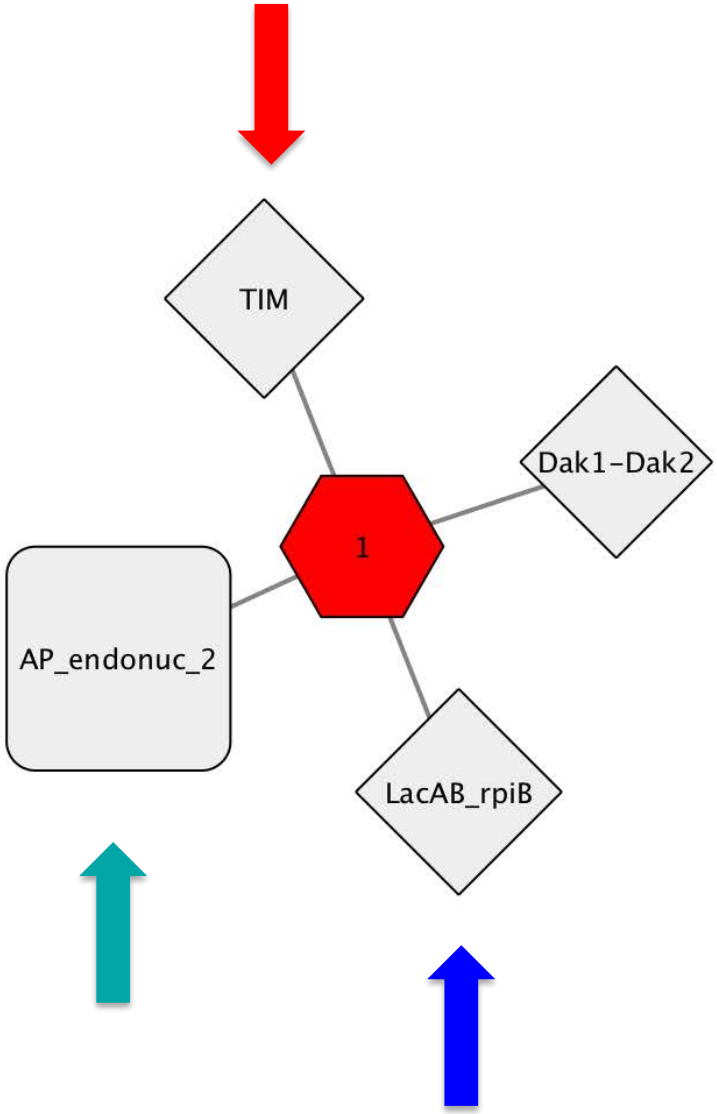


500 total sequences
Alignment score 80
70% ID

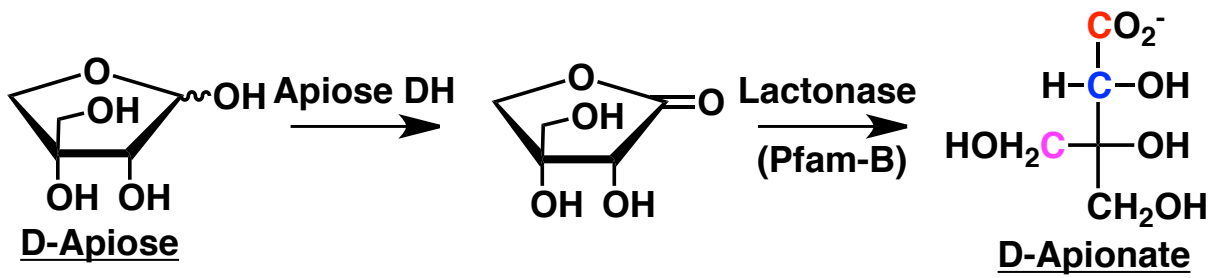
GNN: two (or more) oxidative pathways ?

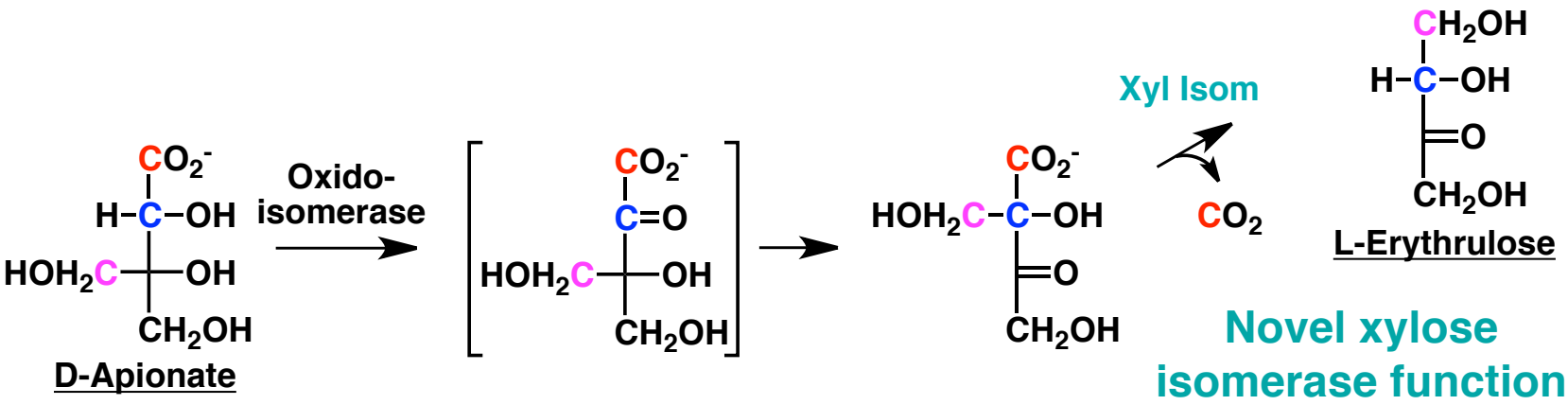


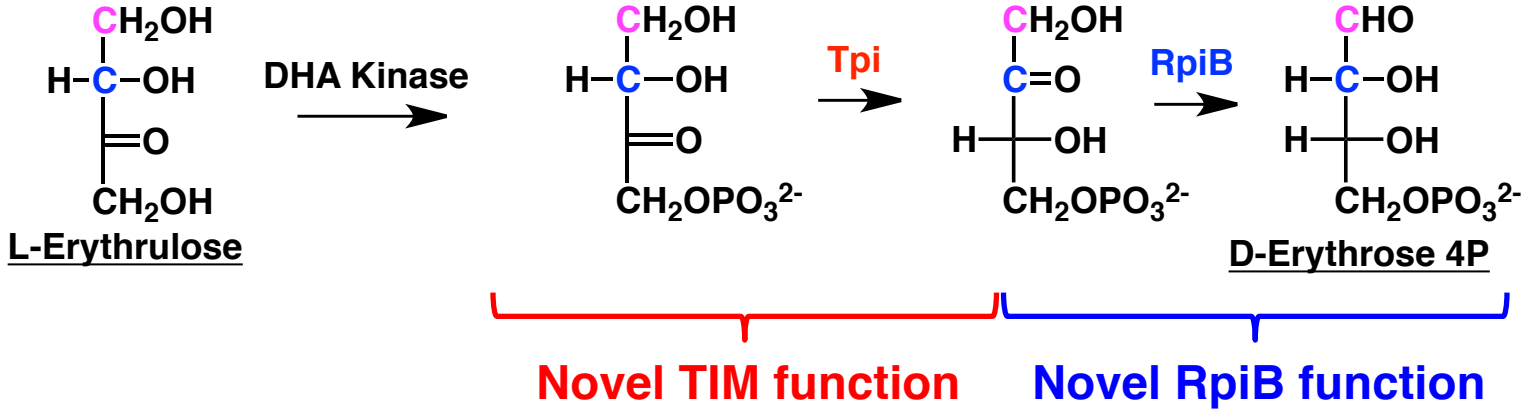
AP endonuclease, TIM, and RpiB families ?



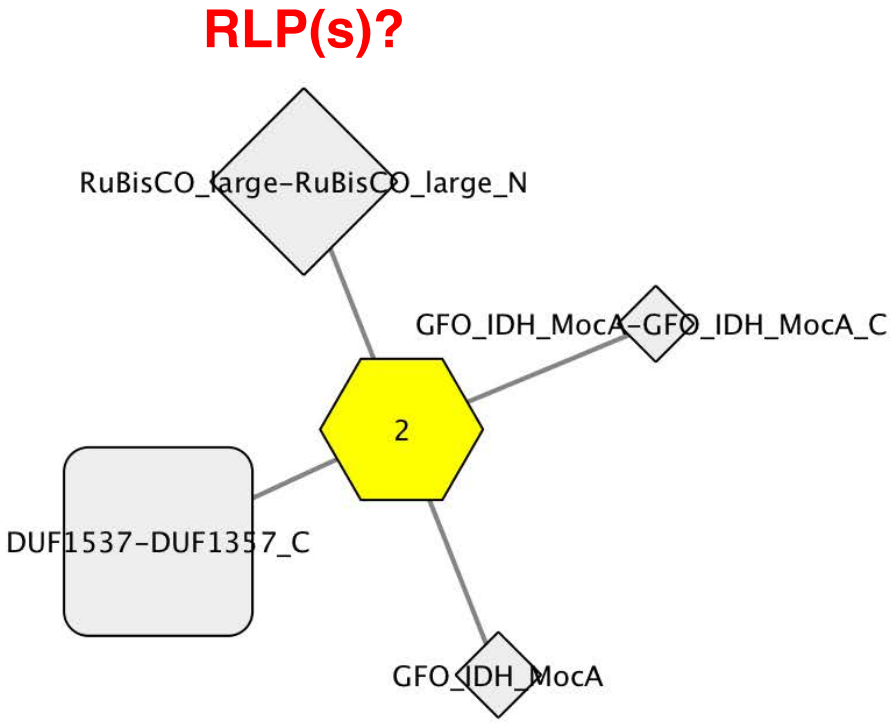
D-Apiose to D-erythrose 4P





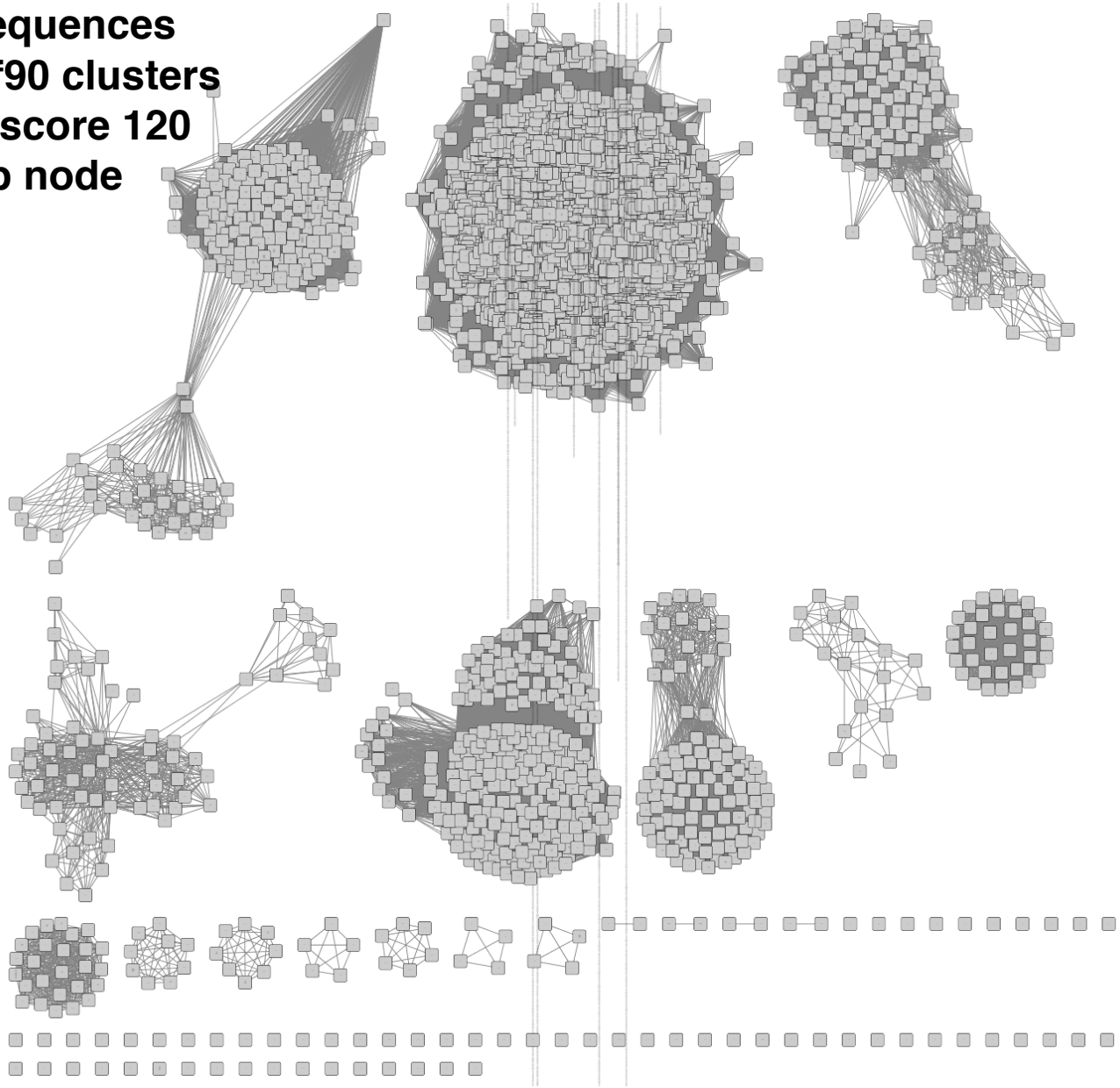


Functions of RuBisCO-Like Proteins (RLPs) ?



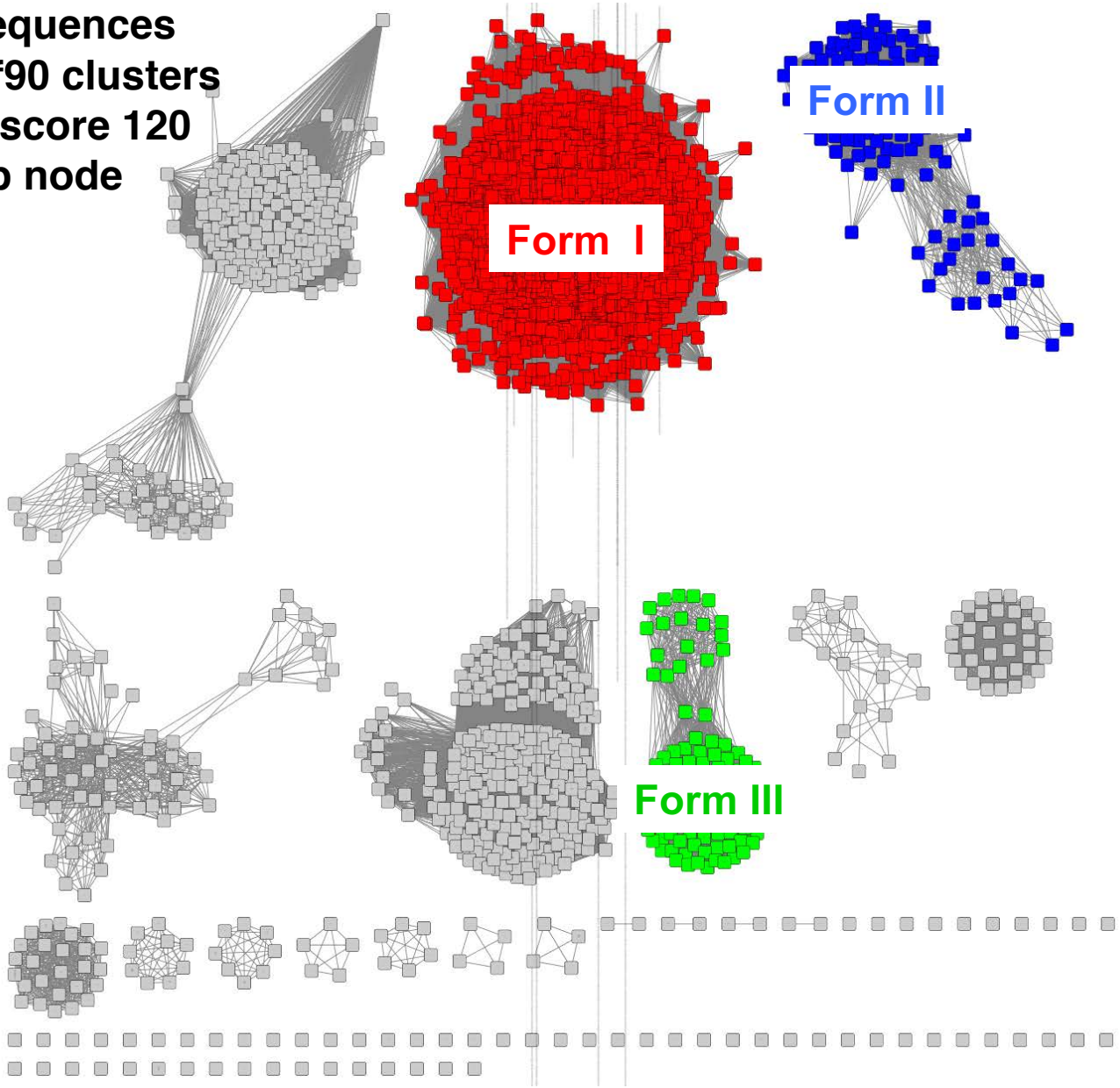
RuBisCO superfamily (PF00016)

136,880 sequences
4,980 UniRef90 clusters
Alignment score 120
95% rep node



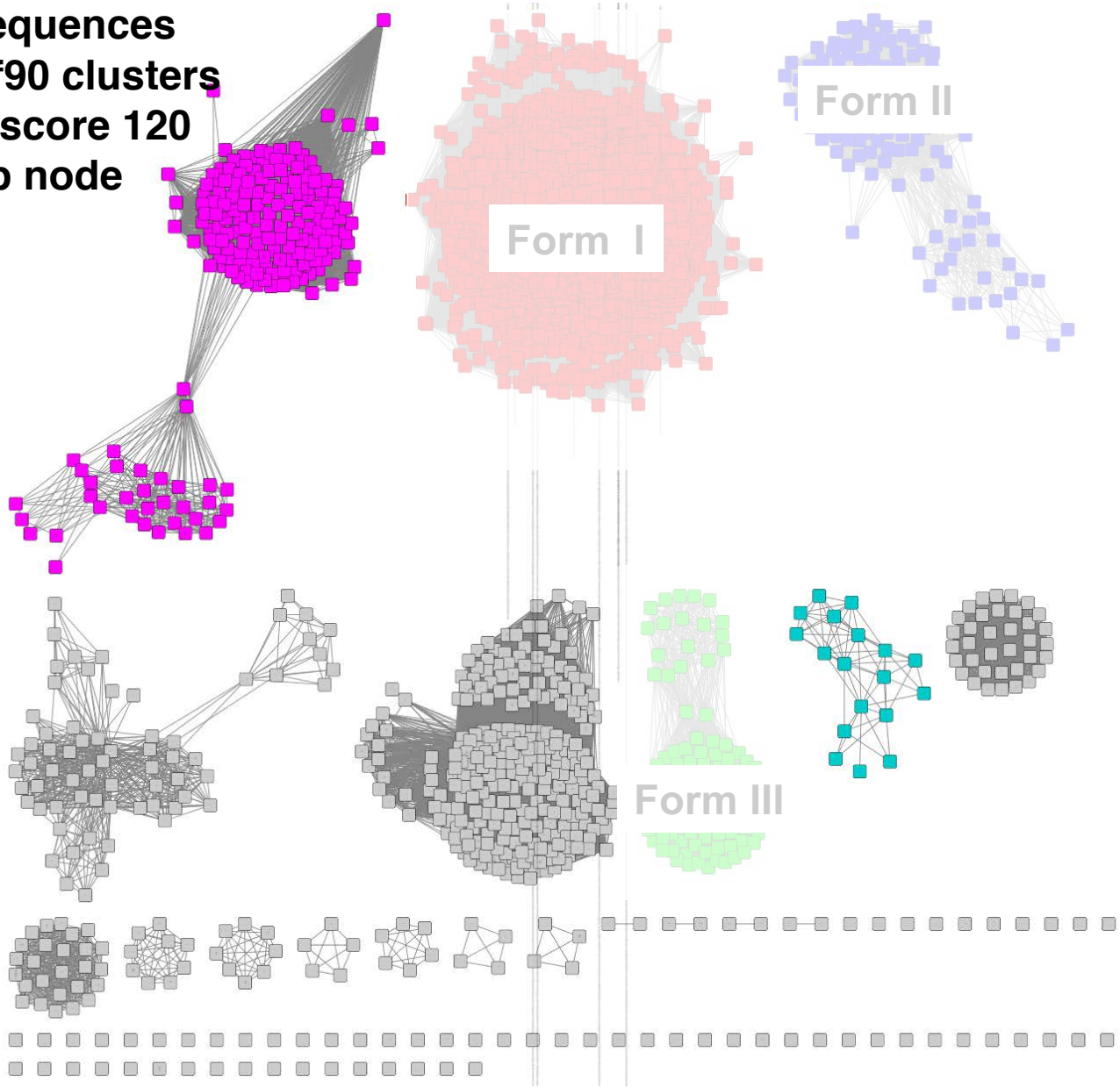
RuBisCO superfamily: three forms of RuBisCO

136,880 sequences
4,980 UniRef90 clusters
Alignment score 120
95% rep node



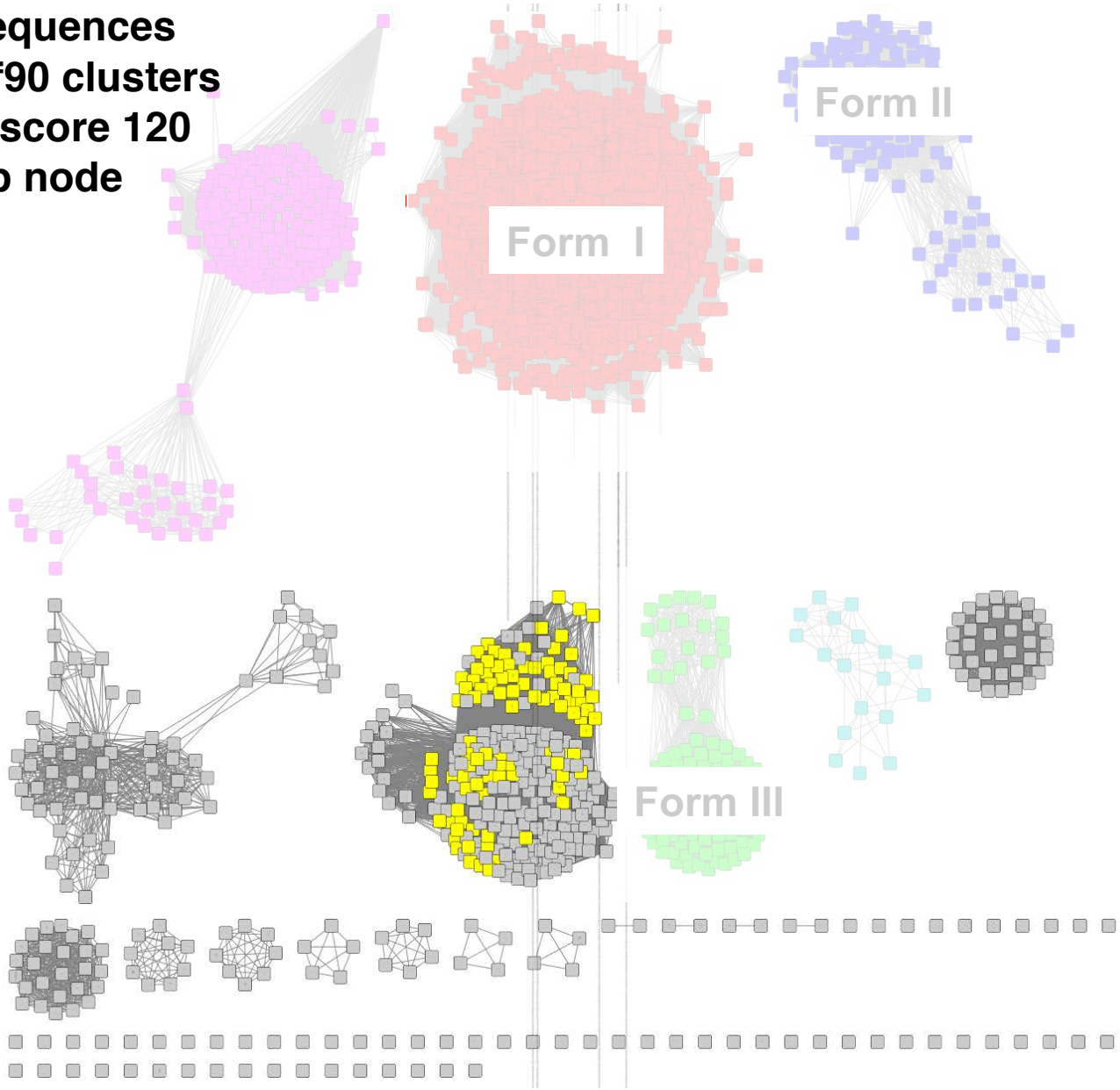
RLPs: Not RuBisCOs, enolate intermediates

136,880 sequences
4,980 UniRef90 clusters
Alignment score 120
95% rep node

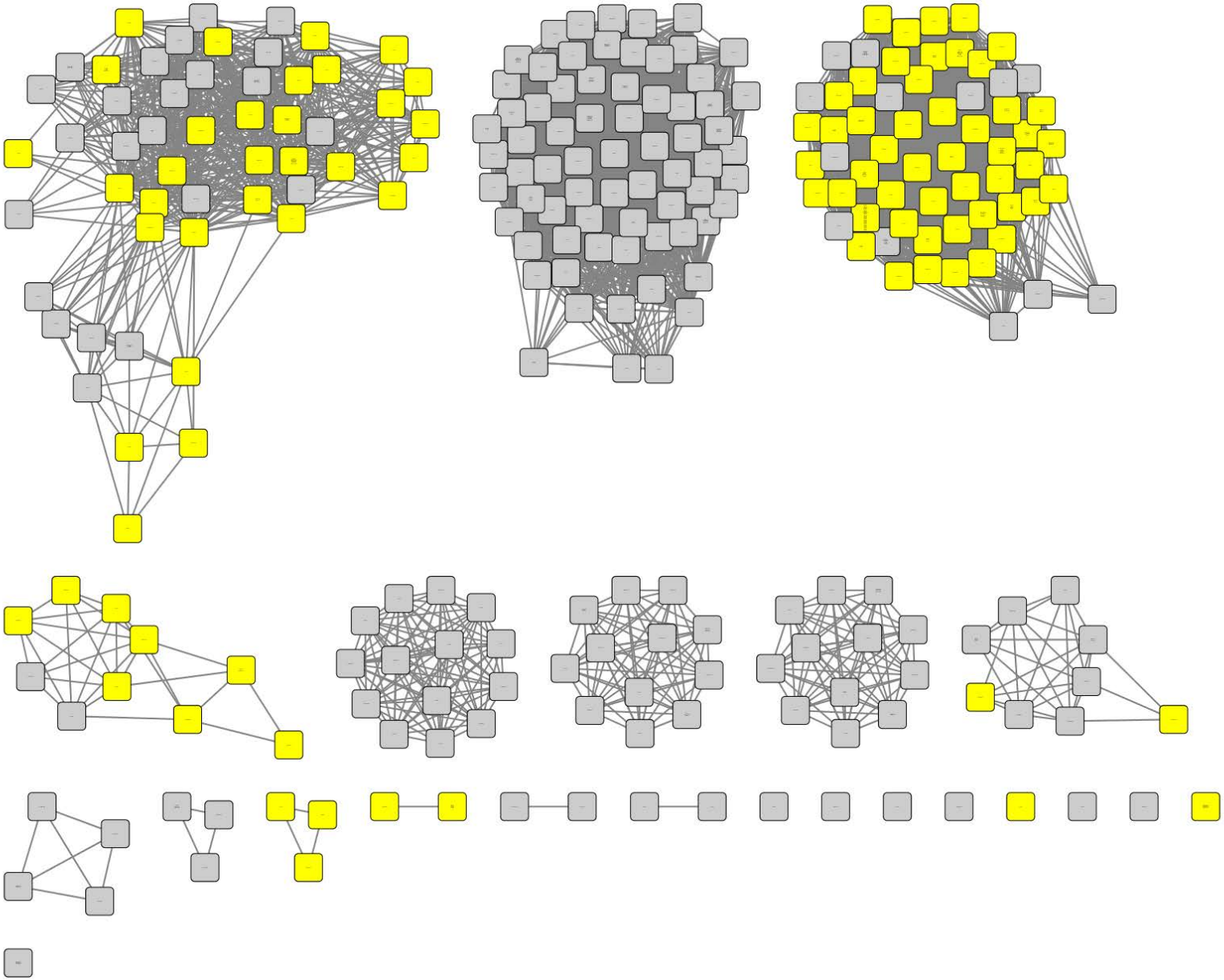


PF16896-RLP Neighbors

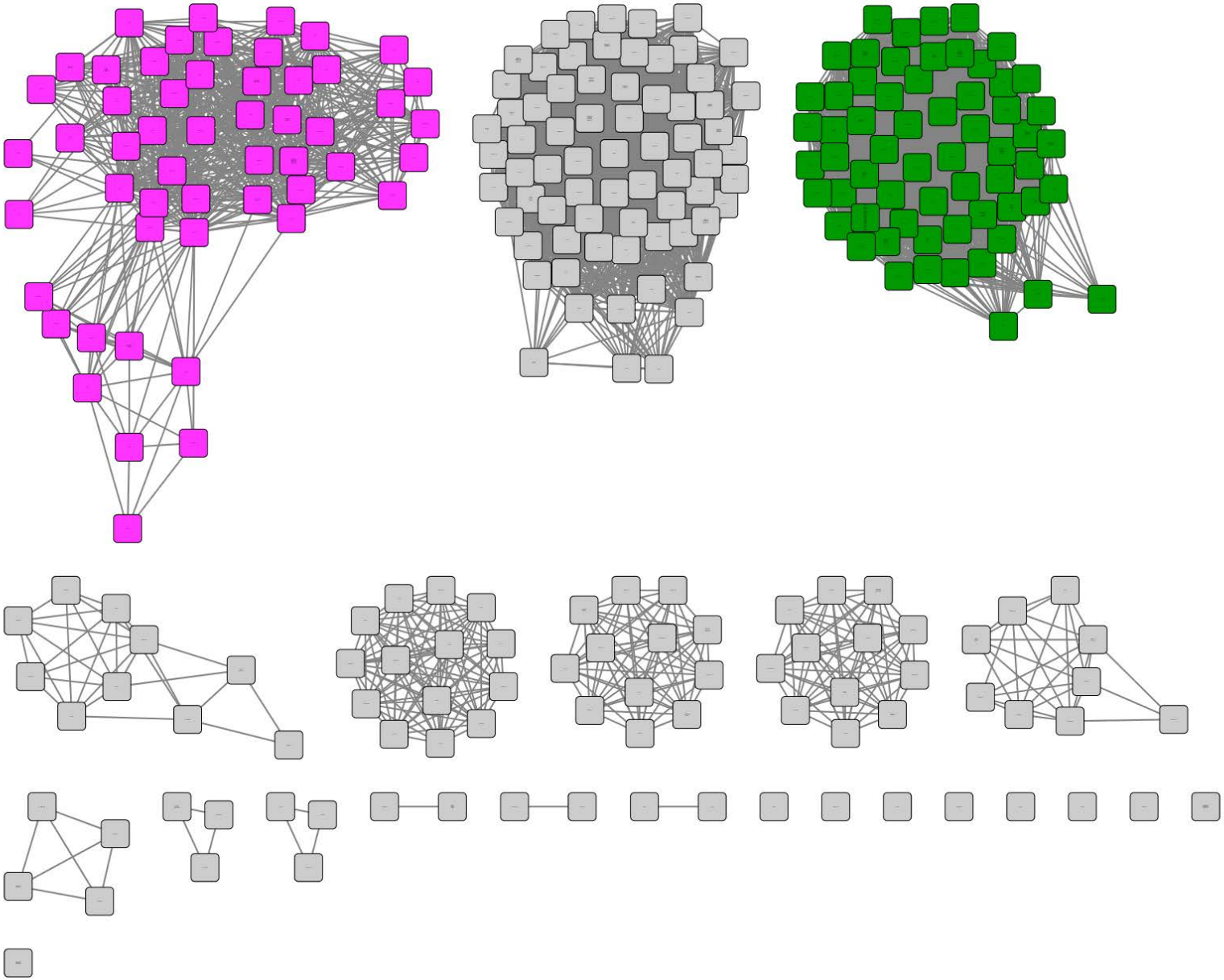
136,880 sequences
4,980 UniRef90 clusters
Alignment score 120
95% rep node



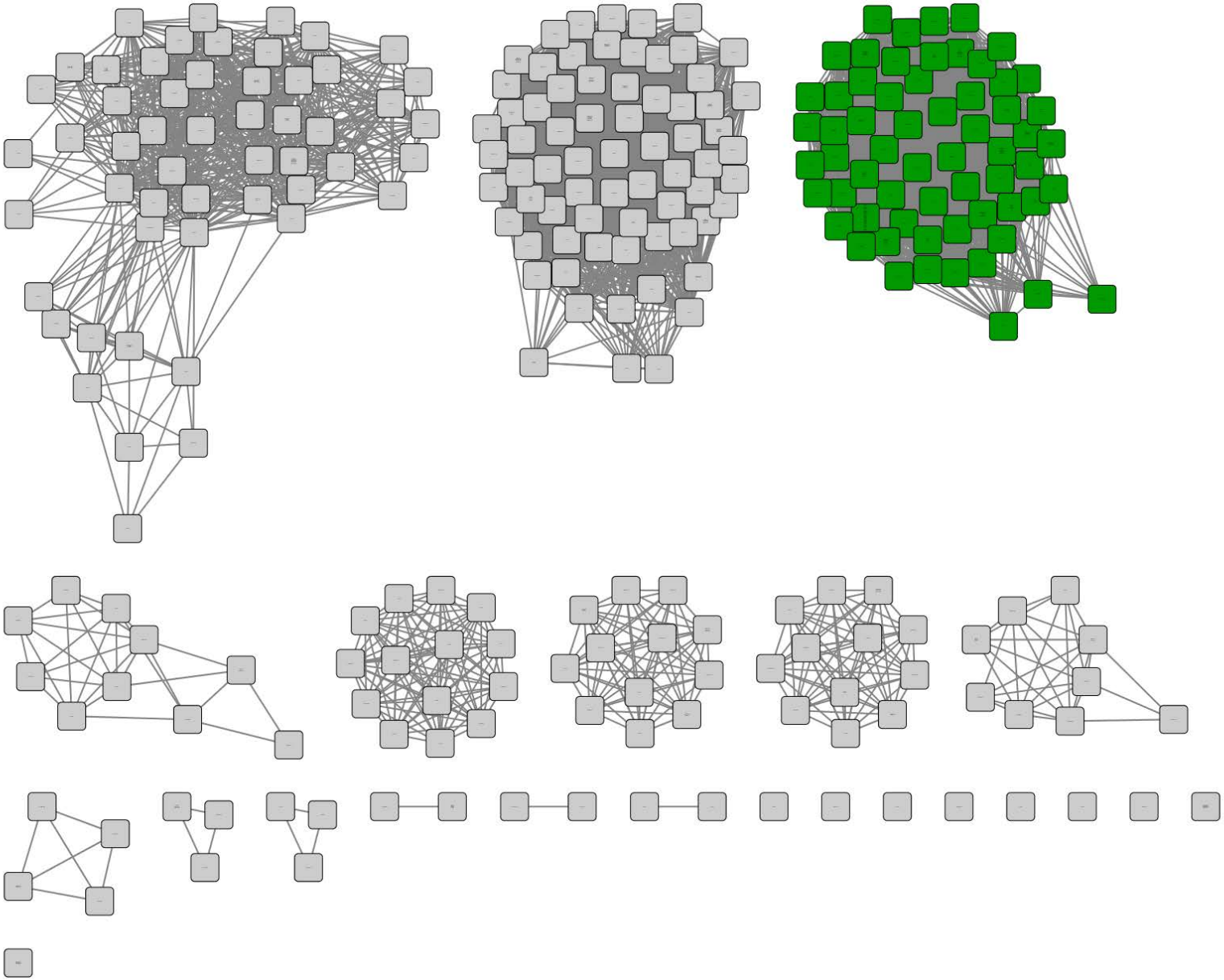
PF16896-RLP Neighbors: alignment score 145



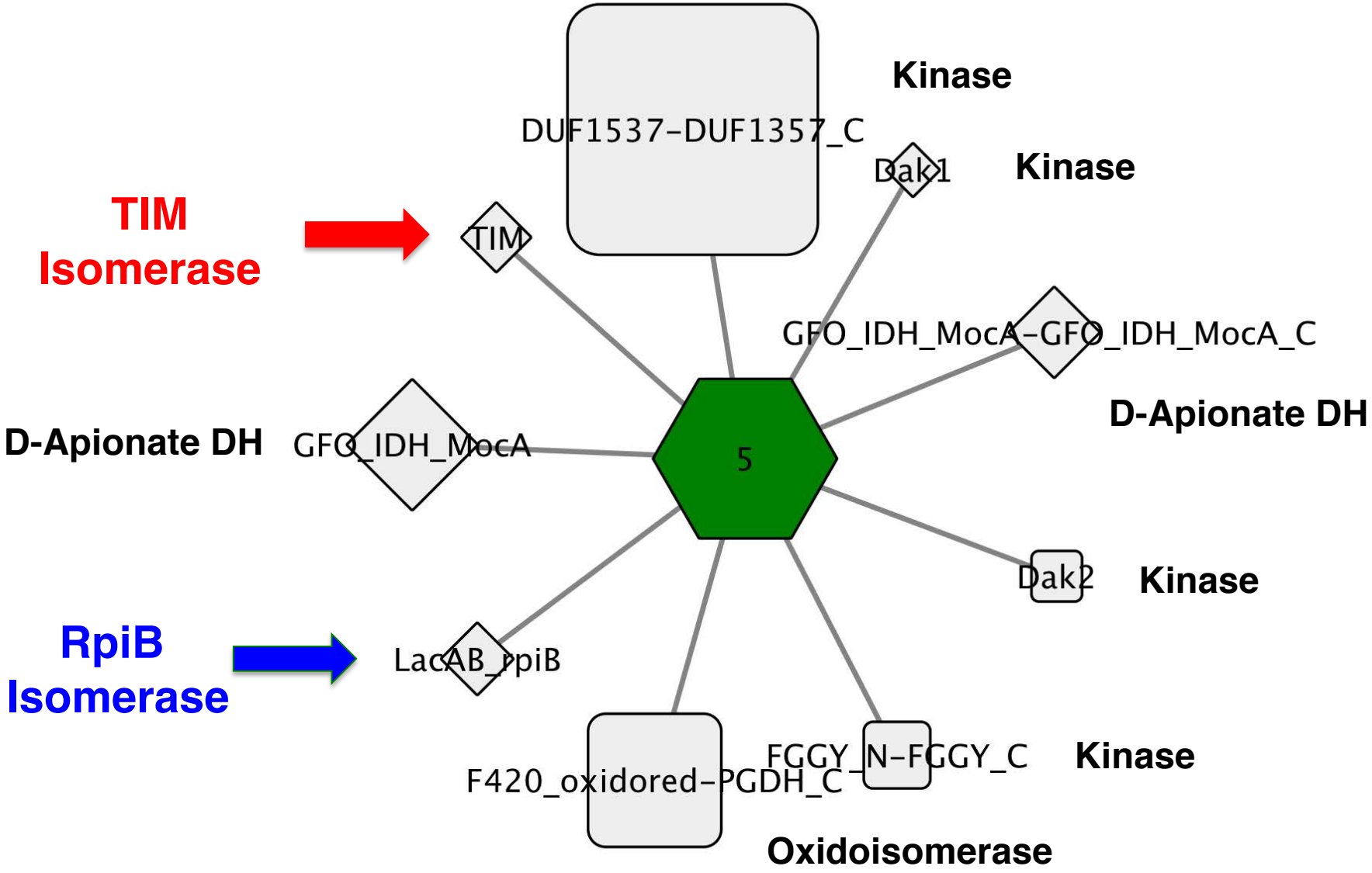
Two major families of PF16896-proximal RLPs



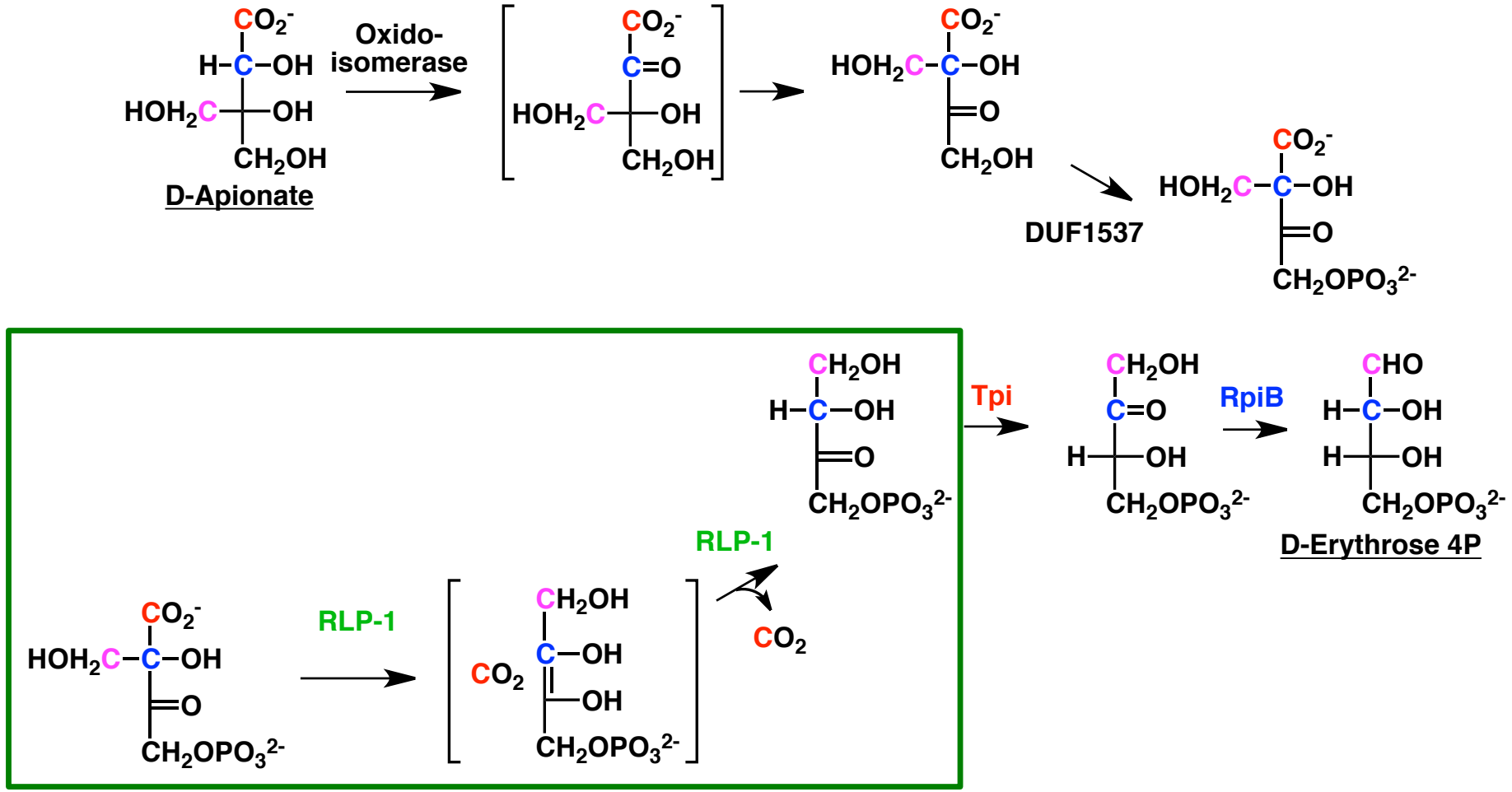
Two major families of PF16896-proximal RLPs



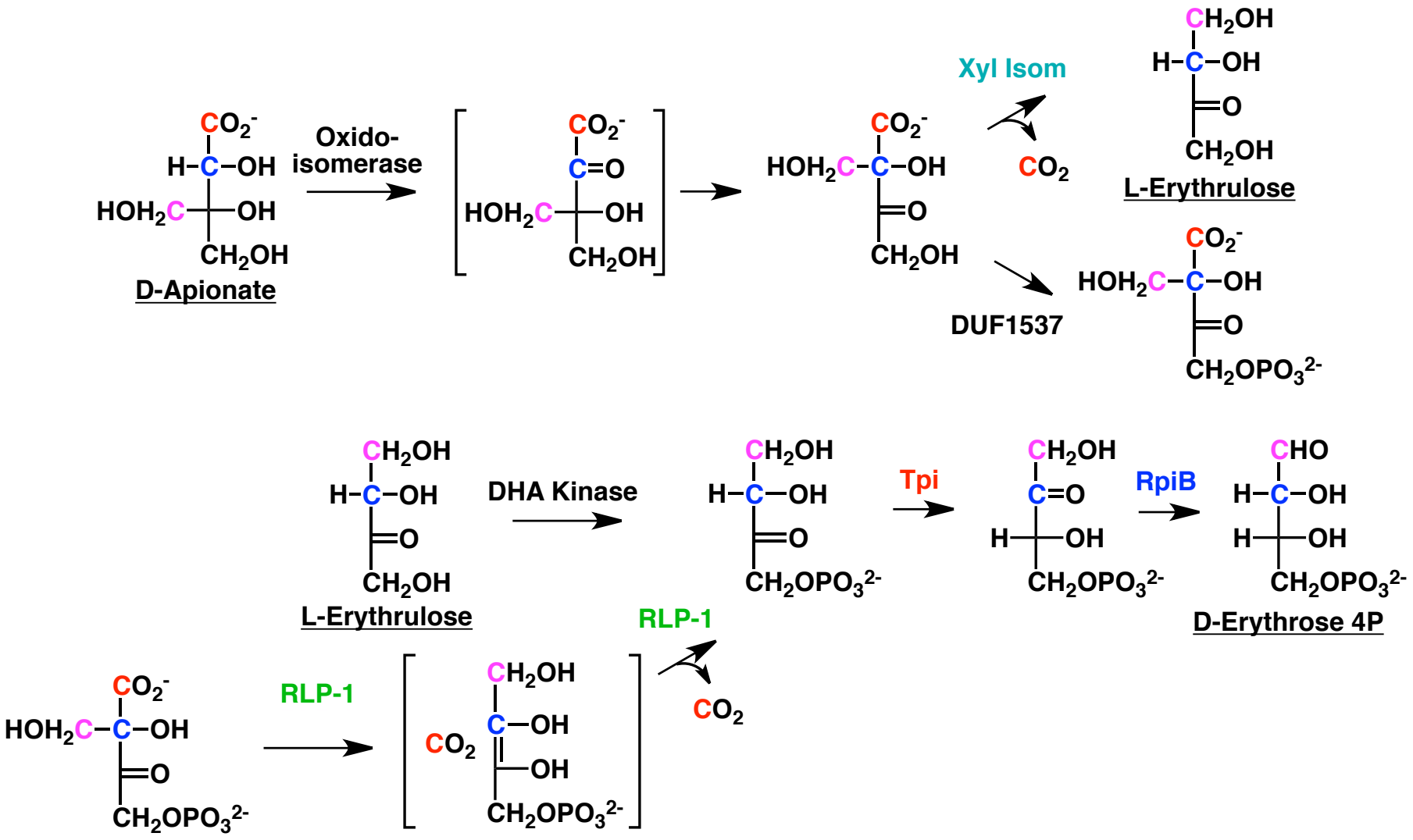
RLP-1: orthologues of TIM and RpiB isomerases in "xylose isomerase" pathway



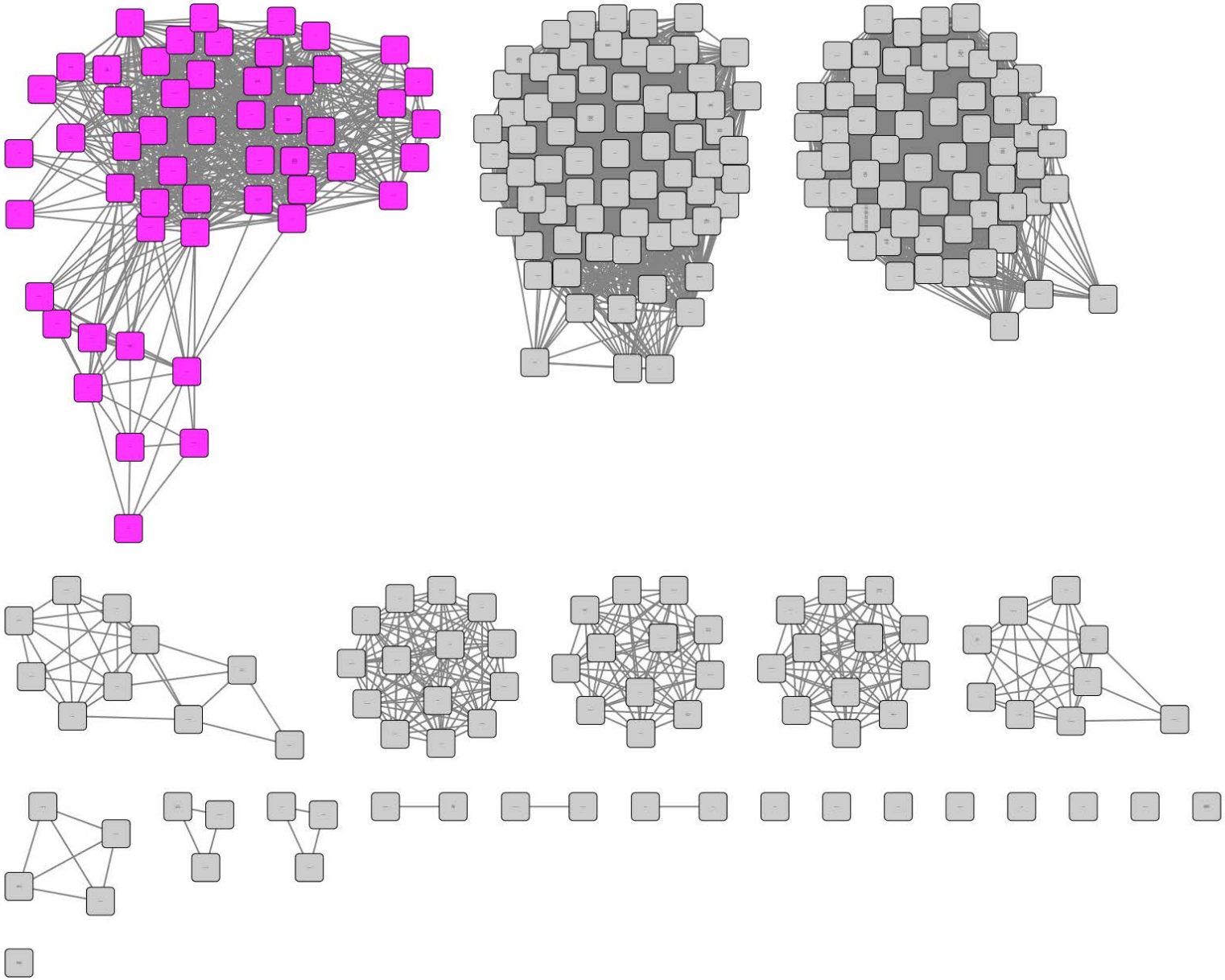
RLP-1: decarboxylase in the RuBisCO superfamily



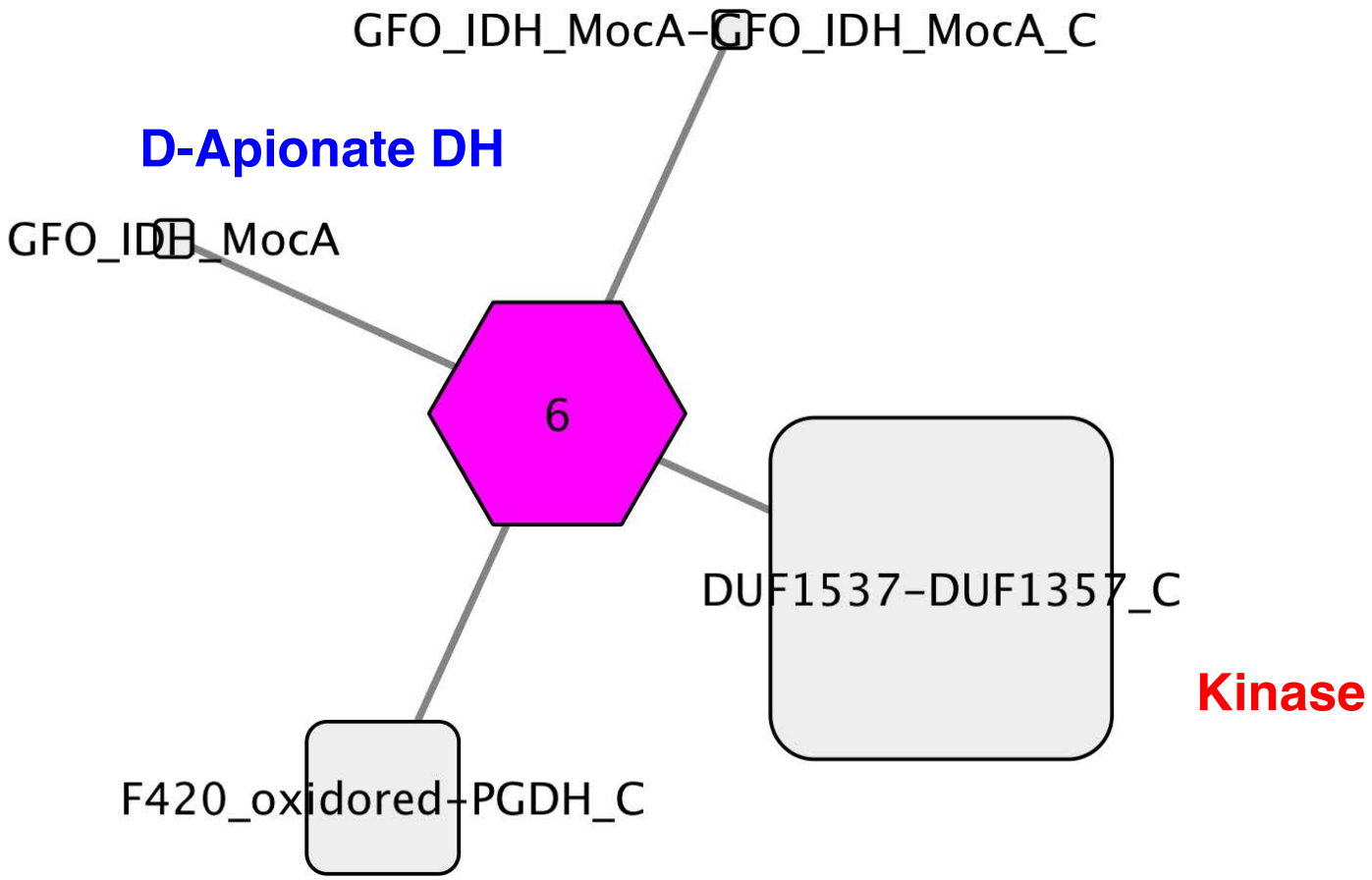
Two pathway variants to D-erythrose 4P (distinct decarboxylase families)



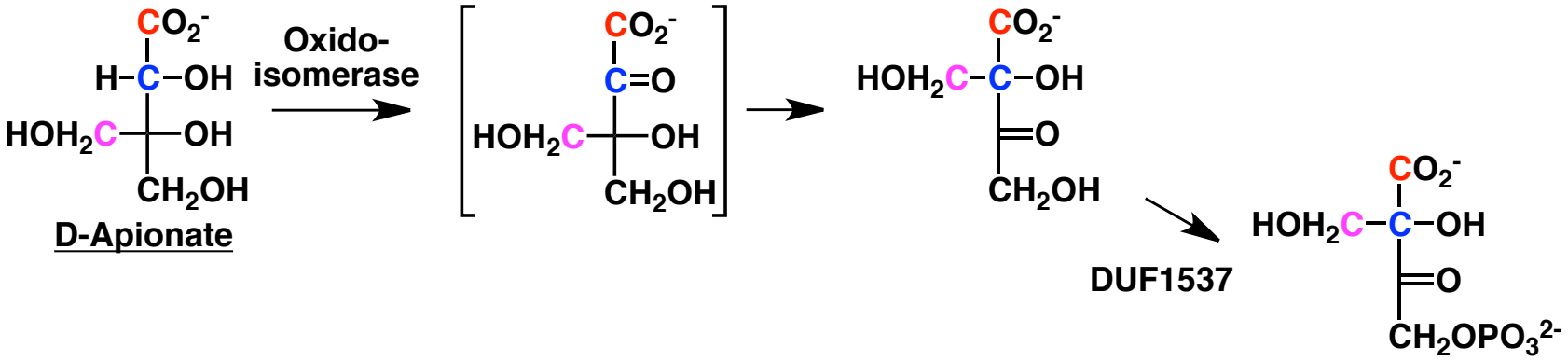
Two major families of PF16896-proximal RLPs

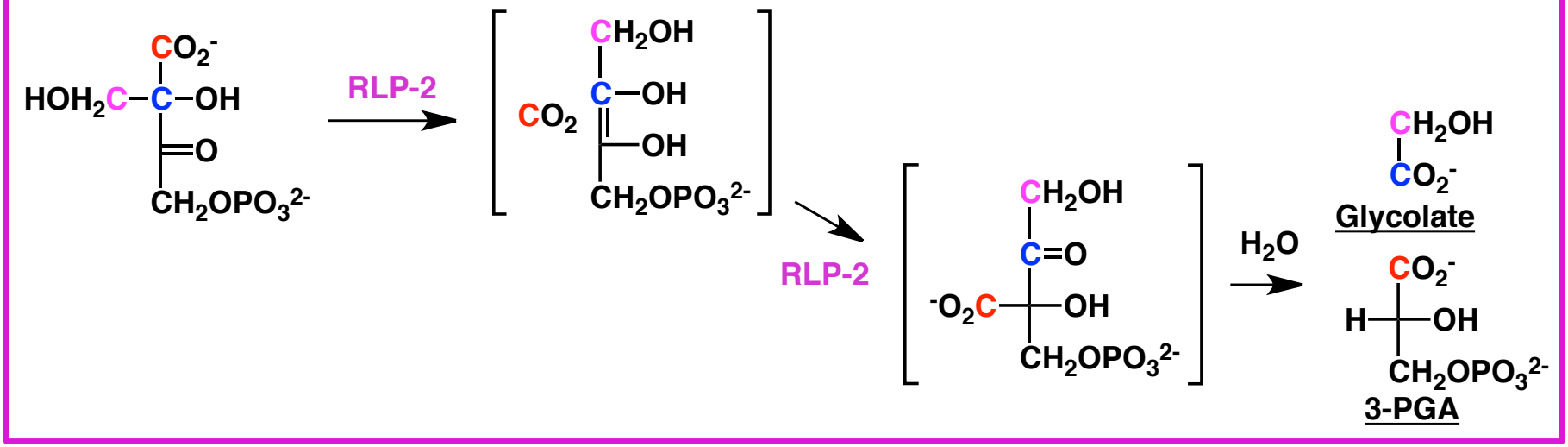


RLP-2: kinase, dehydrogenase, and oxidoisomerase

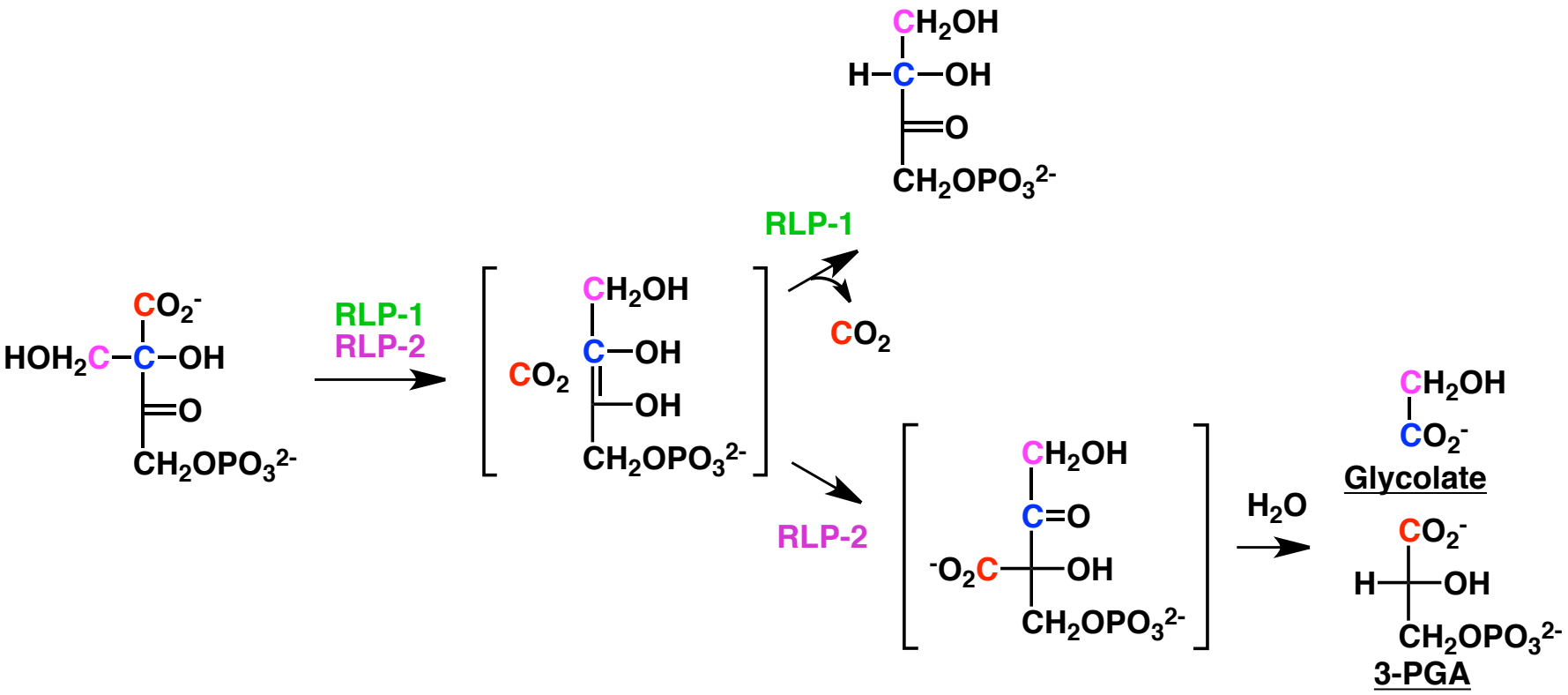


RLP-2: "transcarboxylase" in the RuBisCO superfamily

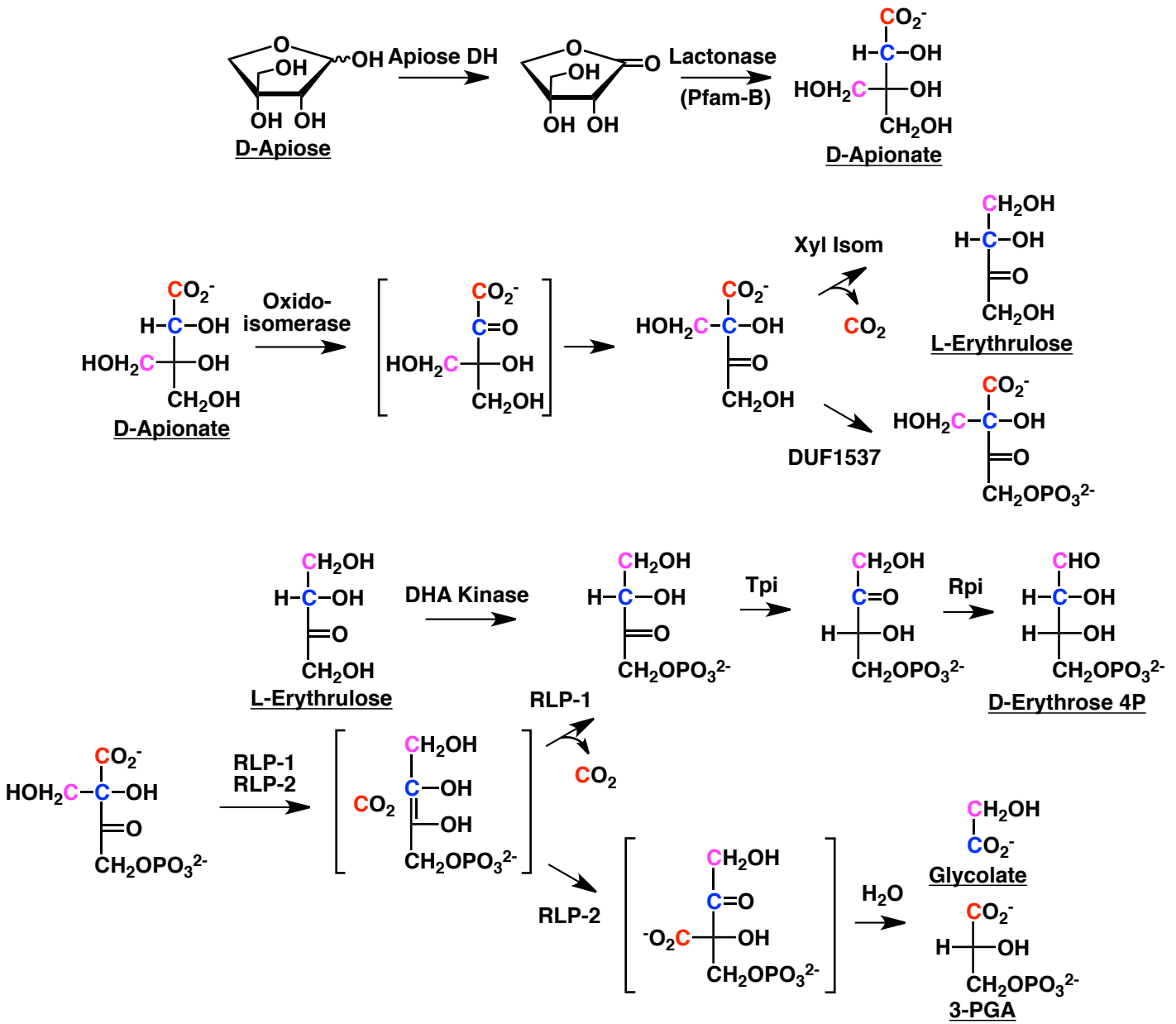




RLP-1/RLP-2: conserved decarboxylation, divergent fates of common enediolate intermediate



Three SBPs → four different pathways in >1300 species



Today

1. Target selection using sequence similarity networks: EFI-EST web tool

2. Pathway discovery using genome neighborhood networks: EFI-GNT web tool

3. Example: using SSNs and GNNs to discover catabolic pathways for D-apiose

4. **Target prioritization using chemically guided functional profiling: CGFP-ShortBRED beta tool**

Chemically guided functional profiling

RESEARCH ARTICLE

MICROBIOTA

A prominent glycyl radical enzyme in human gut microbiomes metabolizes *trans*-4-hydroxy-L-proline

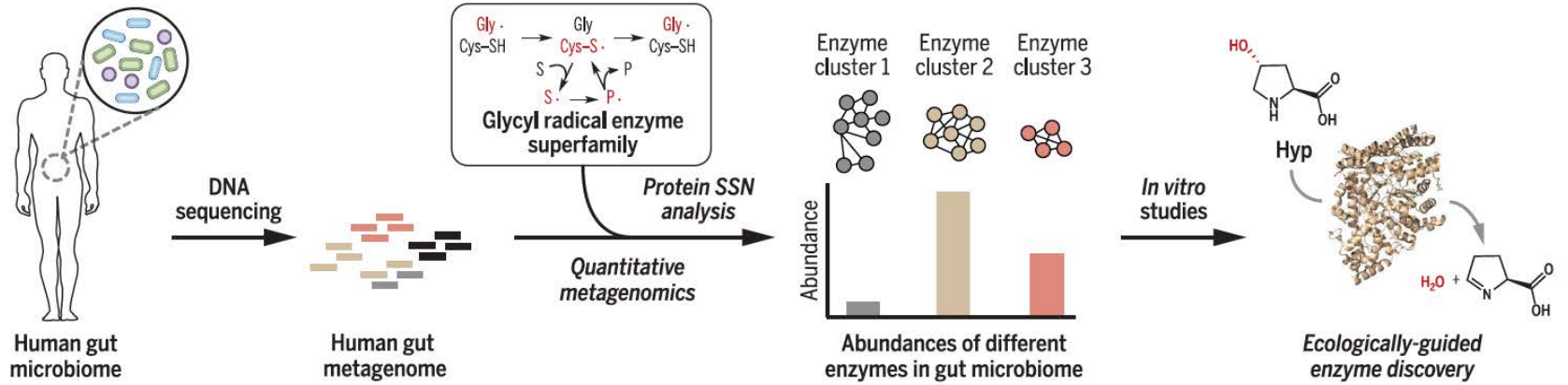
B. J. Levin,[1*] Y. Y. Huang,[1*] S. C. Peck,[1] Y. Wei,[2] A. Martínez-del Campo,[1] J. A. Marks,[1] E. A. Franzosa,[3,4] C. Huttenhower,[3,4] E. P. Balskus[1,4]†

The human microbiome encodes vast numbers of uncharacterized enzymes, limiting our functional understanding of this community and its effects on host health and disease. By incorporating information about enzymatic chemistry into quantitative metagenomics, we determined the abundance and distribution of individual members of the glycyl radical enzyme superfamily among the microbiomes of healthy humans. We identified many uncharacterized family members, including a universally distributed enzyme that enables commensal gut microbes and human pathogens to dehydrate *trans*-4-hydroxy-L-proline, the product of the most abundant human posttranslational modification. This "chemically guided functional profiling" workflow can therefore use ecological context to facilitate the discovery of enzymes in microbial communities.

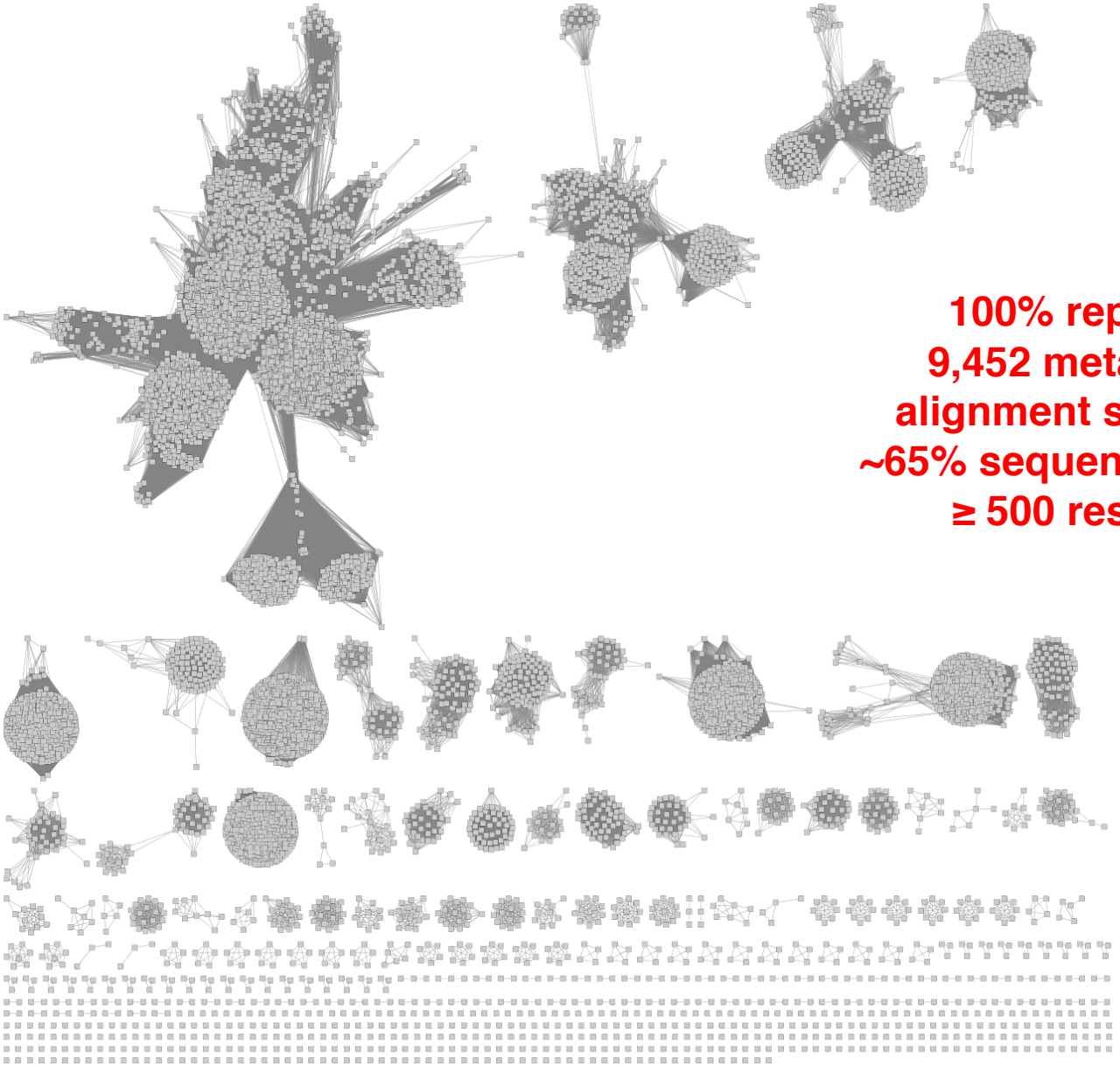
Levin *et al.*, *Science* **355**, 595 (2017) 10 February 2017

Chemically guided functional profiling

Chemically guided functional profiling enables enzyme discovery in microbiomes. Combining protein sequence similarity network (SSN) analysis with quantitative metagenomics reveals the abundance of both characterized and uncharacterized members of enzyme superfamilies. An analysis of glycyl radical enzymes in healthy human microbiomes facilitated the discovery of *trans*-4-hydroxy-L-proline dehydratase, a universally distributed but previously unknown gut microbial enzyme.

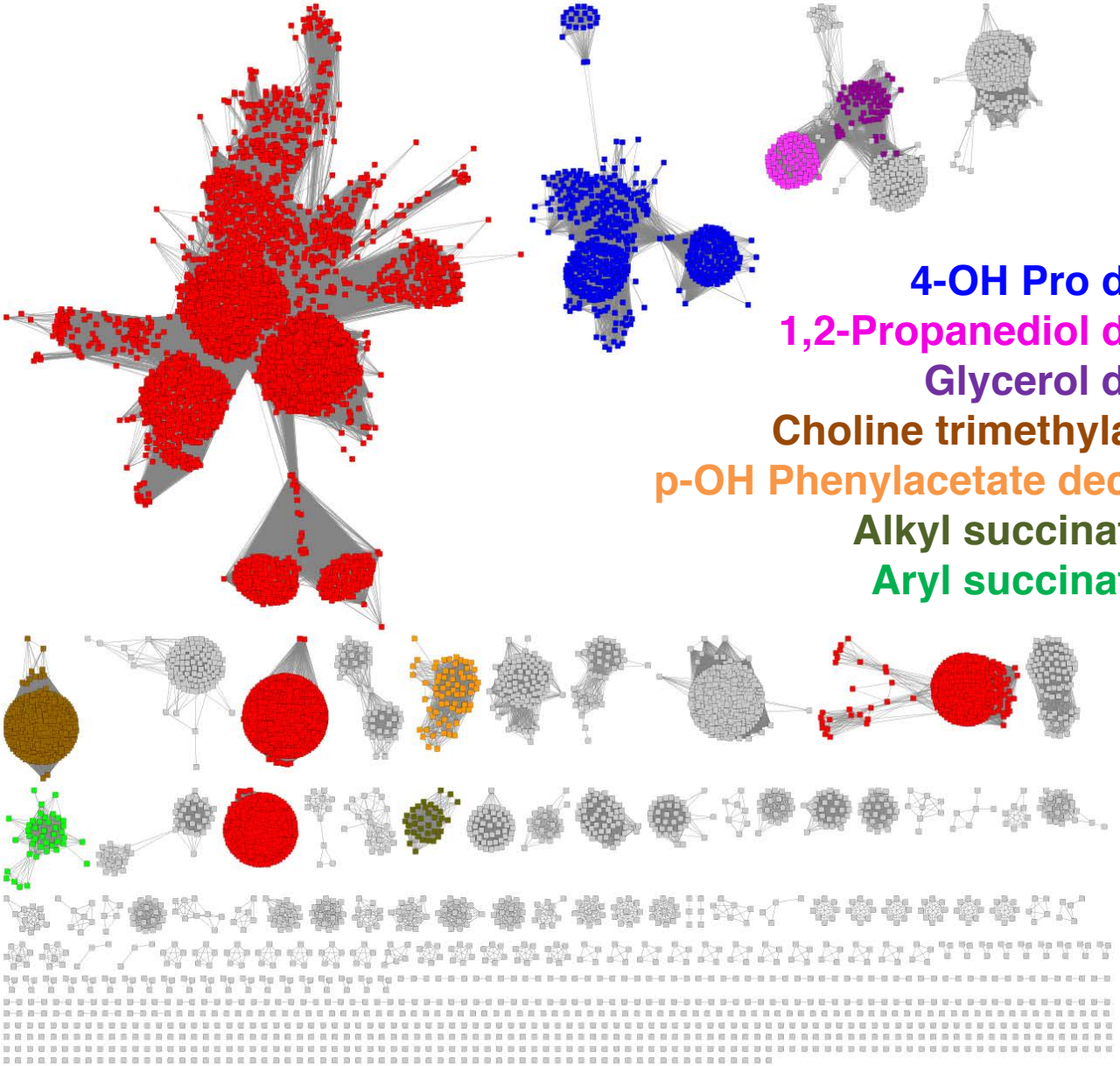


Glycyl radical enzyme (GRE) superfamily: 14,229 sequence



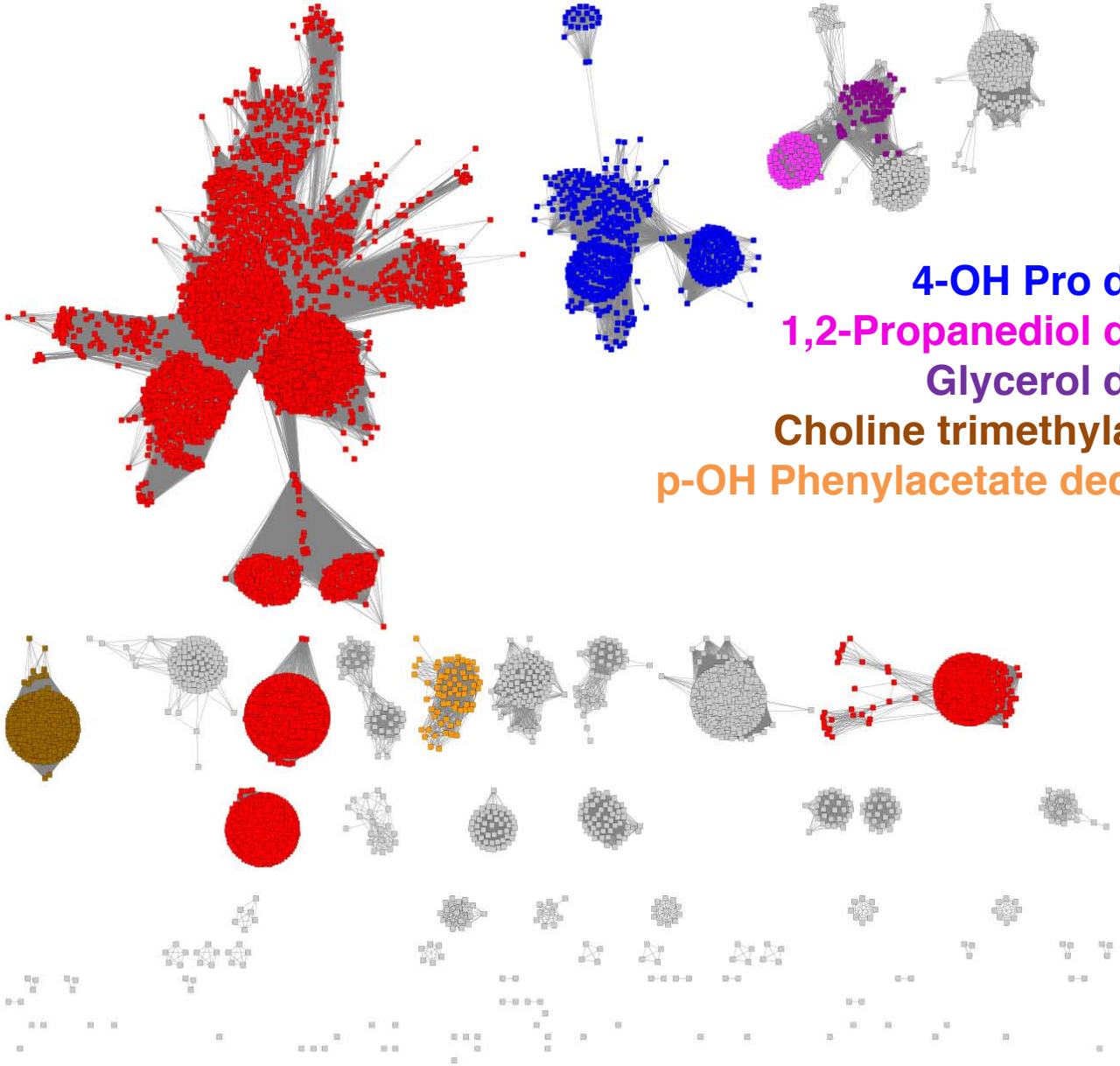
100% rep node
9,452 metanodes
alignment score 300
~65% sequence identity
≥ 500 residues

Clusters with known functions

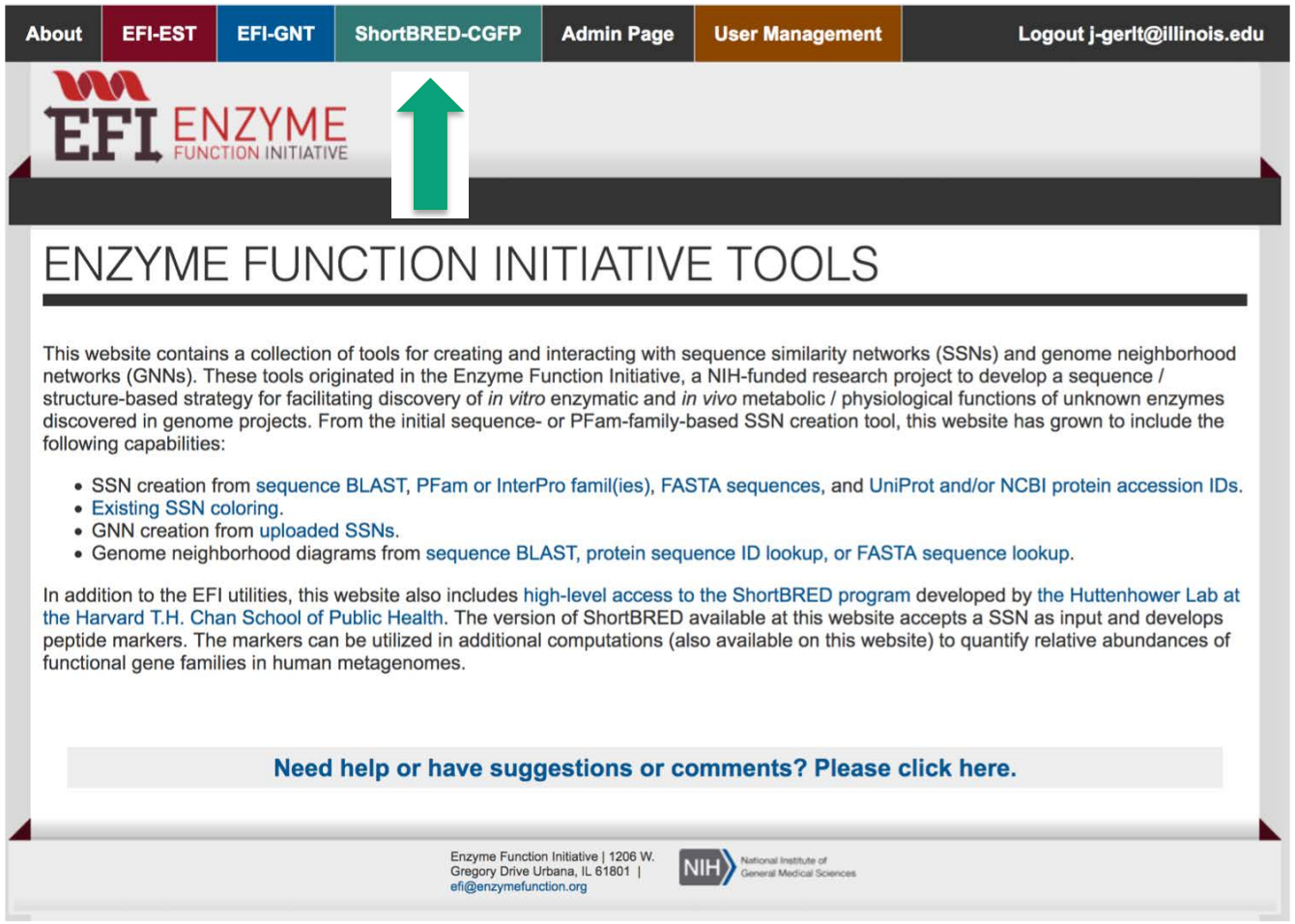


PFL
4-OH Pro dehydratase
1,2-Propanediol dehydratase
Glycerol dehydratase
Choline trimethylamine lyase
p-OH Phenylacetate decarboxylase
Alkyl succinate synthase
Aryl succinate synthase

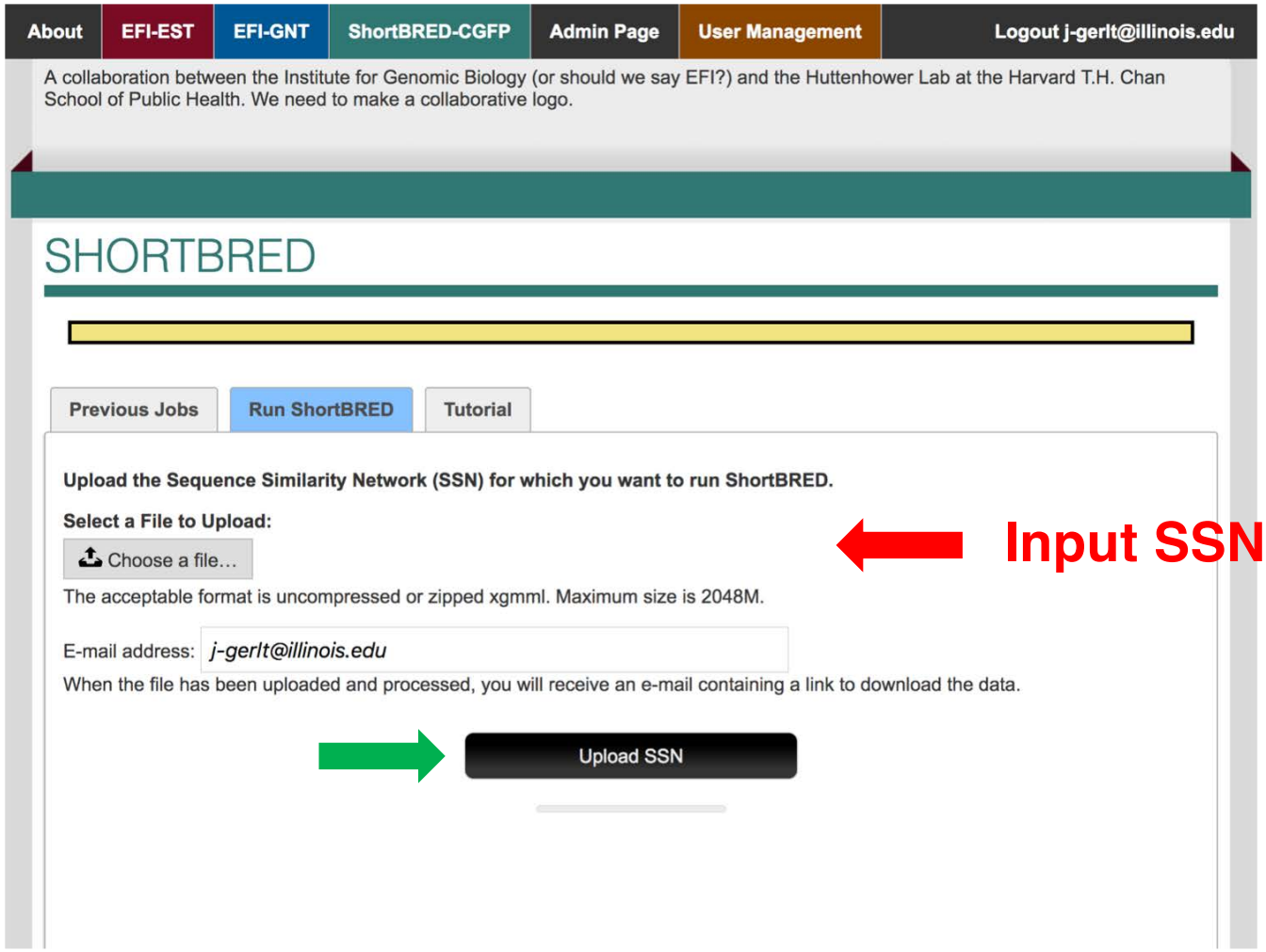
Clusters with orthologues in human gut metagenomes



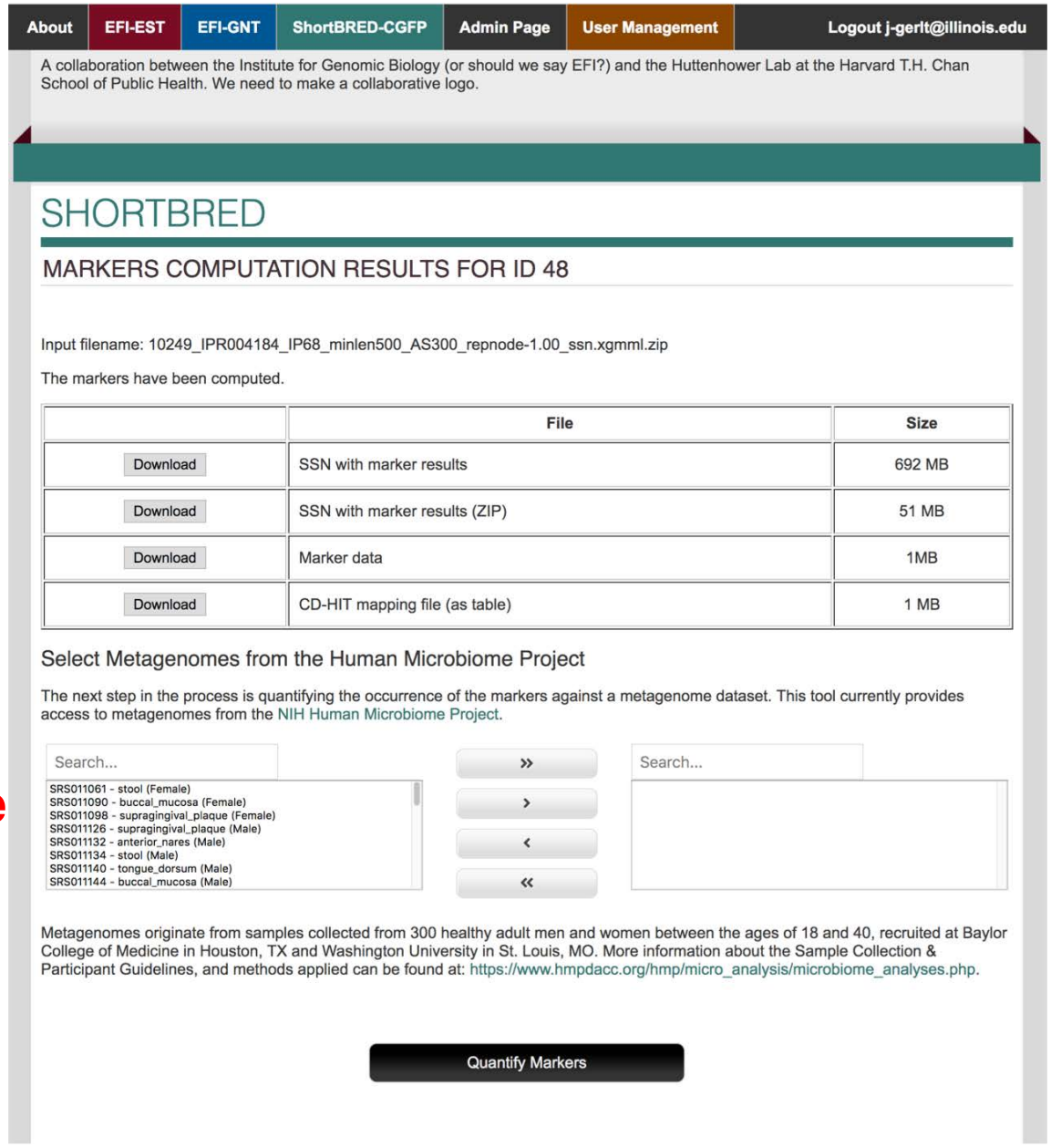
Beta tool: ShortBRED-CGFP



ShortBRED: Identify family-specific markers

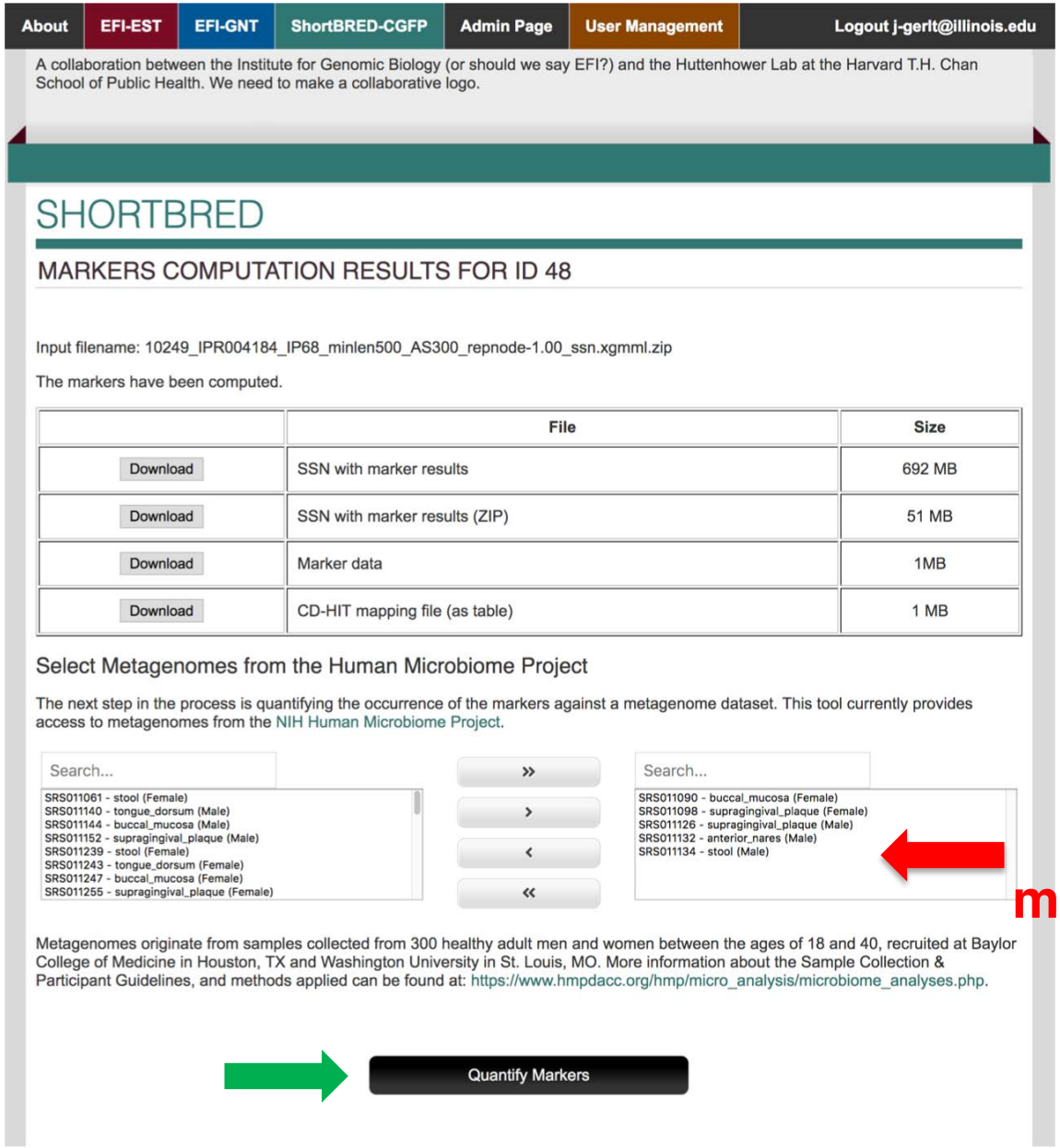


ShortBRED: Quantitate metagenome abundance



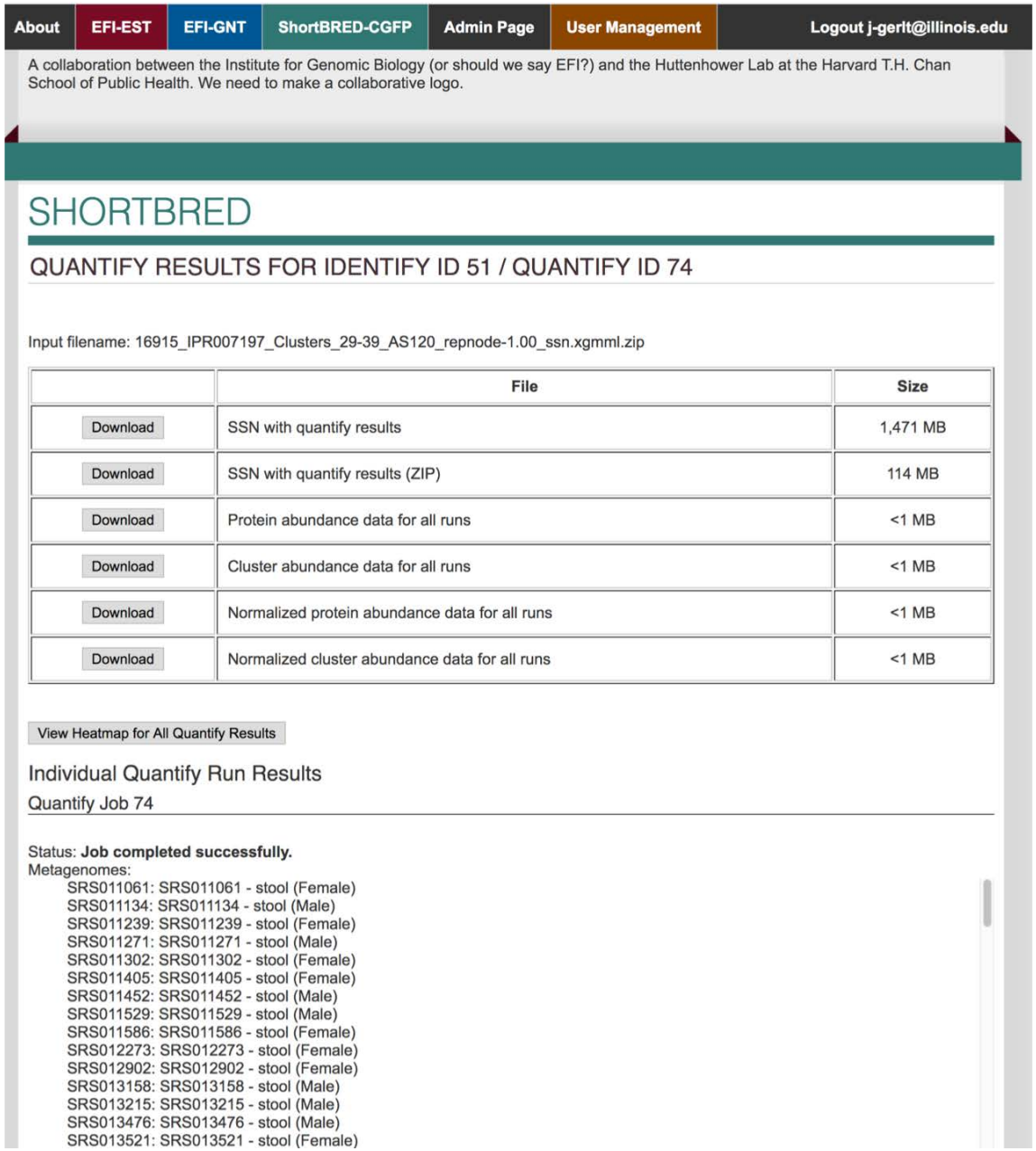
Metagenome library

ShortBRED: Quantitate metagenome abundance



Selected metagenomes

ShortBRED: Output files



Genomic enzymology strategy for functional assignment

1. **Explore sequence function-space in a protein family, segregating family into isofunctional clusters using known functions (SwissProt, literature)**

   Sequence similarity networks (SSNs) using EFI-EST

2. **Predict novel functions of isofunctional clusters in metabolic pathways using genome context (types of reactions enabled by proximal enzymes)**

   Genome neighborhood networks (GNNs) using EFI-GNT

3. **Prioritize clusters for functional characterization by (metagenome) abundance**

   Chemically guided functional profiling (CGFP)

Acknowledgements

Illinois

Nils Oberg
Remi Zallot
Dan Davidson
David Slater
Michael Carter
Xinshuai Zhang
Hua Huang
Brian San Francisco
Tyler Stack
John Cronan
John Gerlt

UCSF

Chakrapani Kalyanaraman
Sara Calhoun
Magdalena Korczynska
Matt Jacobson
Andrej Sali
Brian Shoichet
Patricia Babbitt

Harvard

Emily Balskus
Curtis Huttenhower

Albert Einstein

Matt Vetting
Jeff Bonanno
Agni Ghosh
Scott Garforth
Tyler Grove
Art Arcinis
Anthony Gizzi
Steven Almo

NIH U54GM093342
NIH P01GM118303

EFI-EST: efi.igb.illinois.edu/efi-est/
EFI-GNT: efi.igb.illinois.edu/efi-gnt/

Requirements

- **Registered account for EFI**
- http://efi.igb.illinois.edu

- **Java SE Runtime Environment 8**
- http://www.oracle.com/technetwork/java/javase/downloads/jre8-downloads-2133155.html

- **Cytoscape 3.3**
- Form experience, this version (not the latest) is the best for our needs
- http://chianti.ucsd.edu/cytoscape-3.3.0/
- http://www.cytoscape.org/download_old_versions.html

- Within Cytoscape, install:
- **BridgeDB** from the App manager

- All files generated and used are available in the Workshop Dropbox folder
- Direct link: https://www.dropbox.com/sh/zfy0e6jaa8dttz7/AABkrFkcET5j2WdGtANKid-Pa?dl=0