

SUPPLEMENTAL TUTORIALS

EFI-EST, EFI-GNT, and EFI-CGFP: Enzyme Function Initiative (EFI) Web Resource for Genomic Enzymology Tools

Nils Oberg¹, Rémi Zallot^{2,3}, and John A. Gerlt^{1,4,5*}

¹Carl R. Woese Institute for Genomic Biology, ⁴Department of Biochemistry, and ⁵Department of Chemistry, University of Illinois at Urbana-Champaign, 1206 West Gregory Drive, Urbana, Illinois 61801, United States.

²Department of Chemistry, ³Manchester Institute of Biotechnology, The University of Manchester, 131 Princess Street, Manchester M1 7DN, UK

Tutorial 1. Use of the Taxonomy Tool and Filter by Taxonomy: GRE Superfamily

This tutorial provides the details for the jobs described in the **Taxonomy Tool and Filter By Taxonomy: GRE Superfamily** section in the text. The web resource **Training** page (<https://efi.igb.illinois.edu/training/example.php?id=2022>) provides links to 1) the **Taxonomy Taxonomy Tool** jobs used to generate **Taxonomy Sunbursts** and 2) both the **DATASET COMPLETED** and **DOWNLOAD NETWORK FILES** pages for the EFI-EST jobs used for generating the taxonomy category-filtered SSNs.

Taxonomy Sunbursts: Taxonomy Tool Families Option

Complete Sequences and Fragments. The **Taxonomy Sunburst** for complete sequences and fragments was generated from the entries in UniProt Release 2022_04 by entering IPR004184 (pyruvate formate lyase domain) into the **Pfam and/or InterPro Families and/or Pfam clans** box on the **Taxonomy Tool Families Option** page (red arrow), entering the **Job name** (green arrow) and an **E-mail address** (magenta arrow), and clicking “**Submit analysis**” (black arrow).

Previous Jobs Families FASTA Accession IDs

Retrieve taxonomy for families.

The UniProt IDs for family members are identified in UniProtKB with a list of Pfam families, InterPro families, and/or Pfam clans.

Pfam and/or InterPro Families and/or Pfam clans:

IPR004184 

Family	Family Name	Full Size	UniRef90 Size	UniRef50 Size
IPR004184	PFL_dom	25,513	8,545	1,869
Total:		25,513	8,545	1,869
Total Computed:		25,513		

The input format is a single family or comma/space separated list of families. Families should be specified as PFxxxx (five digits), IPRxxxxx (six digits) or CLxxxx (four digits) for Pfam clans.

Filter by Taxonomy can be used to remove UniProt IDs that do not match the specified taxonomy categories.

The remaining UniProt IDs are used to generate the sunburst.

UniRef90 and UniRef50 clusters that contain the UniProt IDs are retrieved from the UniRef90 and UniRef50 databases using the lookup table provided by UniProt/UniRef. Clusters for which the cluster ID (representative sequence) matches the list of families are retained.

The numbers of UniProt IDs and both UniRef90 cluster and UniRef50 cluster IDs are displayed on the sunburst; the UniProt IDs and both UniRef90 cluster and UniRef50 cluster IDs are available for download and/or transfer to the Accession ID option (Option D) of EFI-EST to generate SSNs.

If the lists of UniRef90 or UniRef50 cluster IDs are used to generate SSNs with the Accession IDs option (Option D) of EFI-EST, the lists should (must!) be filtered with the same list of families (Filter by Family) and any specified taxonomy categories (Filter by Taxonomy) used to generate the lists.

This filtering removes the UniRef90 and UniRef50 clusters with cluster IDs ("representative sequences") or internal UniProt IDs that are not members of the specified families or have the selected taxonomy categories.

Fragment Option

UniProt designates a Sequence Status for each member: Complete if the encoding DNA sequence has both start and stop codons; Fragment if the start and/or stop codon is missing. Approximately 10% of the entries in UniProt are fragments.

Fragments: Check to exclude UniProt-defined fragments in the results. (default: off)

For the UniRef90 and UniRef50 databases, clusters are excluded if the cluster ID ("representative sequence") is a fragment.

UniProt IDs in UniRef90 and UniRef50 clusters with complete cluster IDs are removed from the clusters if they are fragments.

Filter by Taxonomy

Length Filter

Job name: (required) 

E-mail address: 

You will be notified by e-mail when your submission has been processed.



Complete Sequences. The **Taxonomy Sunburst** for the complete sequences was generated by entering IPR004184 into the **Pfam and/or InterPro Families and/or Pfam clans** box on the **Taxonomy Tool Families Option** page (red arrow), checking the **Fragments** box in the **Fragment Option** to exclude fragments (blue arrow), entering the **Job name** (green arrow) and an **E-mail address** (magenta arrow), and clicking “**Submit analysis**” (black arrow).

Previous Jobs **Families** FASTA Accession IDs

Retrieve taxonomy for families.

The UniProt IDs for family members are identified in UniProtKB with a list of Pfam families, InterPro families, and/or Pfam clans.

Pfam and/or InterPro Families and/or Pfam clans:

IPR004184 

Family	Family Name	Full Size	UniRef90 Size	UniRef50 Size
IPR004184	PFL_dom	25,513	8,545	1,869
Total:		25,513	8,545	1,869
Total Computed:		25,513		

The input format is a single family or comma/space separated list of families. Families should be specified as PFxxxxx (five digits), IPRxxxxxx (six digits) or CLxxxx (four digits) for Pfam clans.

Filter by Taxonomy can be used to remove UniProt IDs that do not match the specified taxonomy categories.

The remaining UniProt IDs are used to generate the sunburst.

UniRef90 and UniRef50 clusters that contain the UniProt IDs are retrieved from the UniRef90 and UniRef50 databases using the lookup table provided by UniProt/UniRef. Clusters for which the cluster ID (representative sequence) matches the list of families are retained.

The numbers of UniProt IDs and both UniRef90 cluster and UniRef50 cluster IDs are displayed on the sunburst; the UniProt IDs and both UniRef90 cluster and UniRef50 cluster IDs are available for download and/or transfer to the Accession ID option (Option D) of EFI-EST to generate SSNs.

If the lists of UniRef90 or UniRef50 cluster IDs are used to generate SSNs with the Accession IDs option (Option D) of EFI-EST, the lists should (must) be filtered with the same list of families (Filter by Family) and any specified taxonomy categories (Filter by Taxonomy) used to generate the lists.

This filtering removes the UniRef90 and UniRef50 clusters with cluster IDs ("representative sequences") or internal UniProt IDs that are not members of the specified families or have the selected taxonomy categories.

Fragment Option

UniProt designates a Sequence Status for each member: Complete if the encoding DNA sequence has both start and stop codons; Fragment if the start and/or stop codon is missing. Approximately 10% of the entries in UniProt are fragments.

Fragments: Check to exclude UniProt-s in the results. (default: off) 

For the UniRef90 and UniRef50 databases, clusters are excluded if the cluster ID ("representative sequence") is a fragment.

UniProt IDs in UniRef90 and UniRef50 clusters with complete cluster IDs are removed from the clusters if they are fragments.

Filter by Taxonomy

Length Filter

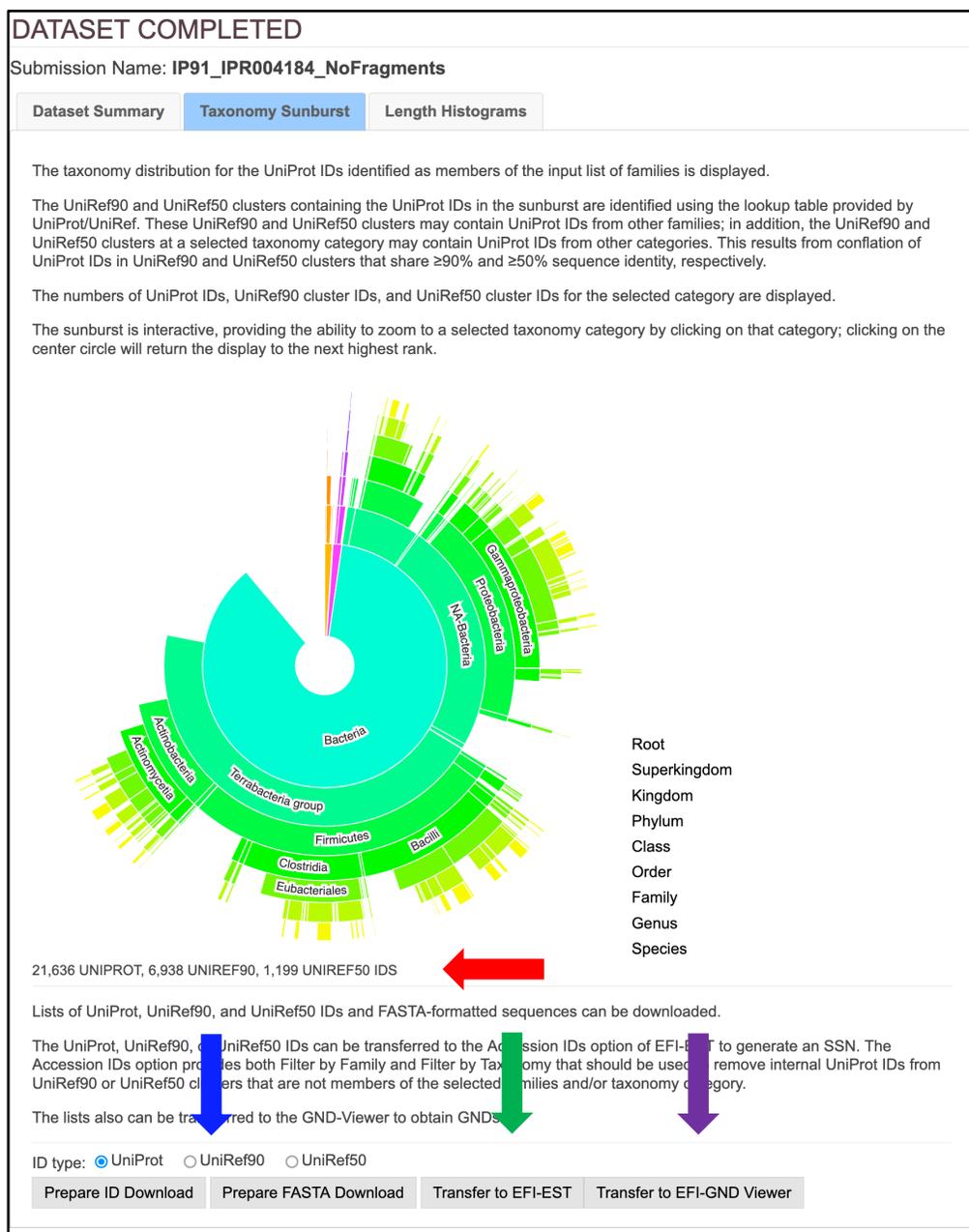
Job name: (required) 

E-mail address: 

You will be notified by e-mail when your submission has been processed.



The results were available on the **DATASET COMPLETED** pages. The **Taxonomy Sunburst** tab (below) is the interactive display that provides the numbers of UniProt, UniRef90 cluster, and UniRef50 cluster IDs (red arrow), downloads for IDs and FASTA sequences (blue arrow), and transfers of IDs to EFI-EST (green arrow) or EFI-GND viewer (magenta arrow).



Complete Sequences, Minimum Length 650 Residues. The Taxonomy Sunburst for complete sequences with a minimum length of 650 residues (“full-length” sequences) was generated by entering IPR004184 into the **Pfam and/or InterPro Families and/or Pfam clans** box on the **Taxonomy Tool Families Option** page (red arrow), checking the **Fragments** box in the **Fragment Option** to exclude fragments (blue arrow), entering 650 for the **Minimum Length** in the **Length Filter** (orange arrow), entering the **Job name** (green arrow) and an **E-mail address** (magenta arrow), and clicking “**Submit analysis**” (black arrow).

Previous Jobs Families FASTA Accession IDs

Retrieve taxonomy for families.

The UniProt IDs for family members are identified in UniProtKB with a list of Pfam families, InterPro families, and/or Pfam clans.

Pfam and/or InterPro Families and/or Pfam clans:

IPR004184 

Family	Family Name	Full Size	UniRef90 Size	UniRef50 Size
IPR004184	PFL_dom	25,513	8,545	1,869
Total:		25,513	8,545	1,869
Total Computed:		25,513		

The input format is a single family or comma/space separated list of families. Families should be specified as PFxxxxx (five digits), IPRxxxxxx (six digits) or CLxxxx (four digits) for Pfam clans.

Filter by Taxonomy can be used to remove UniProt IDs that do not match the specified taxonomy categories.

The remaining UniProt IDs are used to generate the sunburst.

UniRef90 and UniRef50 clusters that contain the UniProt IDs are retrieved from the UniRef90 and UniRef50 databases using the lookup table provided by UniProt/UniRef. Clusters for which the cluster ID (representative sequence) matches the list of families are retained.

The numbers of UniProt IDs and both UniRef90 cluster and UniRef50 cluster IDs are displayed on the sunburst; the UniProt IDs and both UniRef90 cluster and UniRef50 cluster IDs are available for download and/or transfer to the Accession ID option (Option D) of ESI-EST to generate SSNs.

If the lists of UniRef90 or UniRef50 cluster IDs are used to generate SSNs with the Accession IDs option (Option D) of ESI-EST, the lists should (must) be filtered with the same list of families (Filter by Family) and any specified taxonomy categories (Filter by Taxonomy) used to generate the lists.

This filtering removes the UniRef90 and UniRef50 clusters with cluster IDs (“representative sequences”) or internal UniProt IDs that are not members of the specified families or have the selected taxonomy categories.

Fragment Option

UniProt designates a Sequence Status for each member: Complete if the encoding DNA sequence has both start and stop codons; Fragment if the start and/or stop codon is missing. Approximately 10% of the entries in UniProt are fragments.

Fragments: Check to exclude UniProt-defined results. (default: off) 

For the UniRef90 and UniRef50 databases, clusters are excluded if the cluster ID (“representative sequence”) is a fragment.

UniProt IDs in UniRef90 and UniRef50 clusters with complete cluster IDs are removed from the clusters if they are fragments.

Filter by Taxonomy

Length Filter

Minimum Length: 

Maximum Length:

Job name: (required) 

E-mail address: 

You will be notified by e-mail when your submission has been processed.



The results were available on the **DATASET COMPLETED** pages. The **Taxonomy Sunburst** tab (below) is the interactive display that provides the numbers of UniProt, UniRef90 cluster, and UniRef50 cluster IDs (red arrow), downloads for IDs and FASTA sequences (blue arrow), and transfers of IDs to EFI-EST (green arrow) or EFI-GND viewer (magenta arrow).

DATASET COMPLETED

Submission Name: **IP91_IR004184_NoFragments_Minlen650**

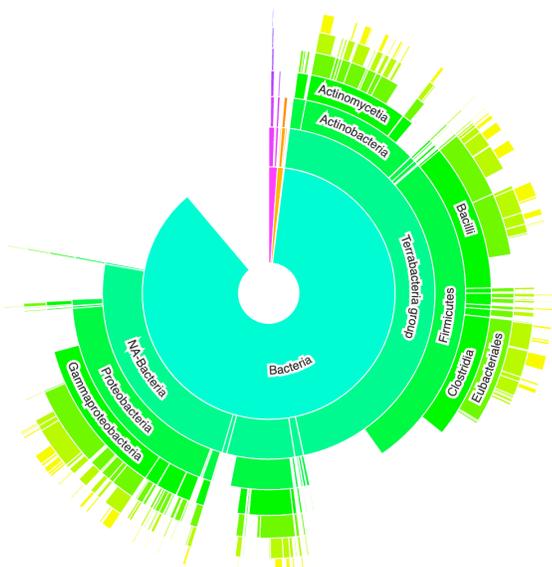
Dataset Summary
Taxonomy Sunburst
Length Histograms

The taxonomy distribution for the UniProt IDs identified as members of the input list of families is displayed.

The UniRef90 and UniRef50 clusters containing the UniProt IDs in the sunburst are identified using the lookup table provided by UniProt/UniRef. These UniRef90 and UniRef50 clusters may contain UniProt IDs from other families; in addition, the UniRef90 and UniRef50 clusters at a selected taxonomy category may contain UniProt IDs from other categories. This results from conflation of UniProt IDs in UniRef90 and UniRef50 clusters that share $\geq 90\%$ and $\geq 50\%$ sequence identity, respectively.

The numbers of UniProt IDs, UniRef90 cluster IDs, and UniRef50 cluster IDs for the selected category are displayed.

The sunburst is interactive, providing the ability to zoom to a selected taxonomy category by clicking on that category; clicking on the center circle will return the display to the next highest rank.



20,089 UNIPROT, 5,802 UNIREF90, 571 UNIREF50 IDS ←

Lists of UniProt, UniRef90, and UniRef50 IDs and FASTA-formatted sequences can be downloaded.

The UniProt, UniRef90, and UniRef50 IDs can be transferred to the **Accession IDs** option of EFI-EST to generate an SSN. The **Accession IDs** option provides both **Filter by Family** and **Filter by Taxonomy** that should be used to remove internal UniProt IDs from UniRef90 or UniRef50 clusters that are not members of the selected families and/or taxonomy category.

The lists also can be transferred to the GND-Viewer to obtain GND.

ID type: UniProt UniRef90 UniRef50

Prepare ID Download
Prepare FASTA Download
Transfer to EFI-EST
Transfer to EFI-GND Viewer

UniProt ID SSN

The UniProt ID SSN was generated for the complete UniProt entries in UniProt Release 2022_04 using the **EFI-EST Families Option** by entering IPR004184 into the **Pfam and/or InterPro Families and/or Pfam clans** box (red arrow), checking the **Fragments** box in the **Fragment Option** to exclude fragments (blue arrow), entering the **Job name** (green arrow) and an **E-mail address** (magenta arrow), and clicking “**Submit analysis**” (black arrow).

Previous Jobs Sequence BLAST **Families** FASTA Accession IDs SSN Utilities

Generate a SSN for a protein family.

The members of the input Pfam families, InterPro families, and/or Pfam clans are selected from the UniProt, UniRef90, or UniRef50 database.

Pfam and/or InterPro Families and/or Pfam clans:

IPR004184 

Use UniRef90 cluster ID sequences instead of UniProt IDs (UniProt is default).

Family	Family Name	Full Size	UniRef90 Size	UniRef50 Size
IPR004184	PFL_dom	25,513	8,545	1,869
Total:		25,513	8,545	1,869
Total Computed:		25,513		

The input format is a single family or comma/space separated list of families. Families should be specified as PFxxxxx (five digits), IPRxxxxxx (six digits) or CLxxxx (four digits) for Pfam clans.

UniRef90 clusters contain UniProt IDs that share $\geq 90\%$ sequence identity and have 80% overlap with the longest sequence in the cluster ("seed sequence"); as a result, the UniProt IDs in the cluster usually are functionally homogeneous, i.e., orthologues. UniRef50 clusters contain UniProt IDs that share $\geq 50\%$ sequence identity and have 80% overlap with the seed sequence; as a result, the UniProt IDs in the cluster often are functionally heterogeneous, e.g., paralogues.

The sequences from the UniRef90 and UniRef50 databases are the UniRef90 and UniRef50 clusters for which the cluster ID ("representative sequence") matches the specified families. The UniProt members in these UniRef90 and UniRef50 clusters that do not match the specified families are removed from the cluster.

Fragment Option

UniProt designates a Sequence Status for each member: Complete if the encoding DNA sequence has both start and stop codons; Fragment if the start and/or stop codon is missing. Approximately 10% of the entries in UniProt are fragments.

Fragments: Check to exclude fragments in the results. (default: off) 

For the UniRef90 and UniRef50 databases, clusters are excluded if the cluster ID ("representative sequence") is a fragment.

UniProt IDs in UniRef90 and UniRef50 clusters with complete cluster IDs are removed from the clusters if they are fragments.

Filter by Taxonomy

Protein Family Size Options

Family Domain Boundary Option

SSN Edge Calculation Option

Job name: (required) 

E-mail address: 

You will be notified by e-mail when your submission has been processed.



The SSN was finalized on the **SSN Finalization** tab of the **DATASET COMPLETED** page using 240 as the **Alignment Score Threshold** that separates the SwissProt-curated functions into different clusters (orange arrow) and 650 residues as the **Minimum** in the **Sequence Length Restriction** to remove truncated sequences (cyan arrow) [1], entering the **Network (SSN) name** (brown arrow), and clicking “**Create SSN**” (black arrow).

DATASET COMPLETED

Submission Name: **IP91_IPR004184_UniProt_NoFragments**

A minimum sequence similarity threshold that specifies the sequence pairs connected by edges is needed to generate the SSN. This threshold also determines the segregation of proteins into clusters. The threshold is applied to the edges in the SSN using the alignment score, an edge node attribute that is a measure of the similarity between sequence pairs.

Dataset Summary | **Taxonomy Sunburst** | **Dataset Analysis** | **SSN Finalization**

This tab is used to specify the minimum "Alignment Score Threshold" (that is a measure of the minimum sequence similarity threshold) for drawing the edges that connect the proteins (nodes) in the SSN.

Alignment Score Threshold: 

This value corresponds to the lower limit for which an edge will be present in the SSN. The alignment score is similar in magnitude to the negative base-10 logarithm of a BLAST e-value.

Sequence Length Restriction Options

Allows restriction of sequences in the generated SSN based on their length. 

Minimum: (default: 0) 

Maximum: (default: 50000)

▸ **Filter by Taxonomy**

▸ **Neighborhood Connectivity**

▸ **Fragment Option**

Network name: This name  is displayed in Cytoscape.

You will be notified by e-mail when the SSN is ready for download.



The **Network Files** tab of the **DOWNLOAD NETWORK FILES** page provided the xgmml file for the **Full (SSN) Network** (red arrow; all UniProt nodes and edges) as well as the xgmml files for **Representative Node Networks** that conflate the UniProt nodes based on percent identity (blue arrow). The xgmml file for the full SSN was downloaded, opened with Cytoscape 3.9.1, and displayed with the yFiles Organic layout to obtain the SNN shown below.

DOWNLOAD NETWORK FILES

Submission Name: IPR004184_UniProt_NoFragments
 Network Name: IPR004184_UniProt_NoFragments_Minlen650_AS240

SSN Overview **Network Files**

Please cite your use of the EFI tools:

Rémi Zallot, Nils Oberg, and John A. Gerit, **The EFI Web Resource for Genomic Enzymology Tools: Leveraging Protein, Genome, and Metagenome Databases to Discover Novel Enzymes and Metabolic Pathways**. *Biochemistry* 2019 58 (41), 4169-4182. <https://doi.org/10.1021/acs.biochem.9b00735>

The panels below provide files for full and representative node SSNs for download with the indicated numbers of nodes and edges. As an approximate guide, SSNs with ~2M edges can be opened with 16 GB RAM, ~5M edges can be opened with 32 GB RAM, ~10M edges can be opened with 64 GB RAM, ~20M edges can be opened with 128 GB RAM, ~40M edges can be opened with 256 GB RAM, and ~120M edges can be opened with 768 GB RAM.

Files may be transferred to the Genome Neighborhood Tool (GNT), the Color SSN utility, the Cluster Analysis utility, or the Neighborhood Connectivity utility.

Full Network ←

Each node in the network represents a single protein sequence.

	# Nodes	# Edges	
Download ZIP	20,089	47,499,276	Transfer To: ▾

Representative Node Networks ←

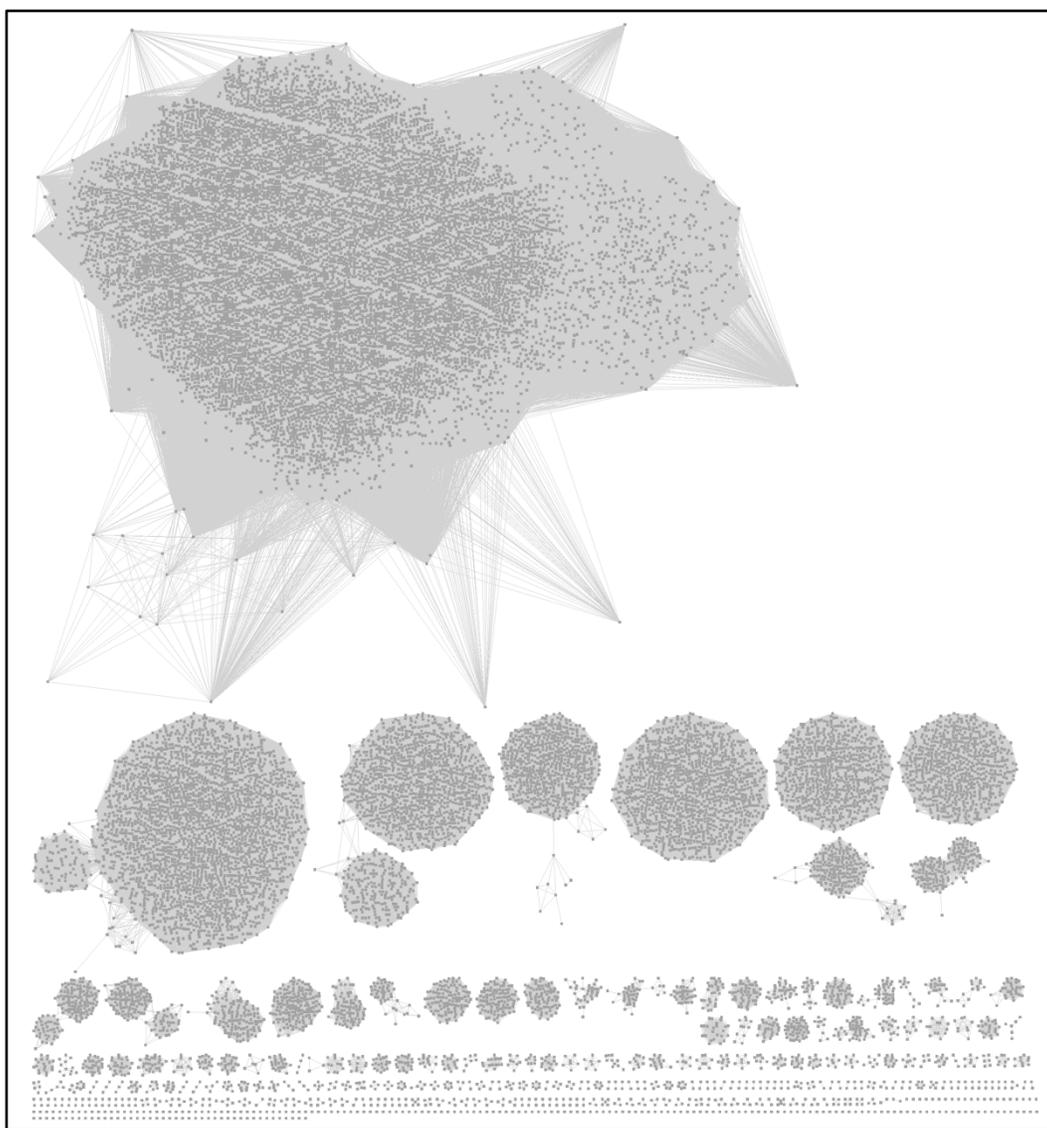
In representative node (RepNode) networks, each node in the network represents a collection of proteins grouped according to percent identity. For example, for a 75% identity RepNode network, all connected sequences that share 75% or more identity are grouped into a single node (meta node). Sequences are collapsed together to reduce the overall number of nodes, making for less complicated networks easier to load in Cytoscape.

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

	% ID	# Nodes	# Edges	
Download ZIP	100	15,986	27,061,467	Transfer To: ▾
Download ZIP	95	7,879	4,912,158	Transfer To: ▾
Download ZIP	90	6,623	2,928,401	Transfer To: ▾
Download ZIP	85	5,685	1,783,730	Transfer To: ▾
Download ZIP	80	4,903	1,112,553	Transfer To: ▾
Download ZIP	75	4,237	636,171	Transfer To: ▾
Download ZIP	70	3,631	325,427	Transfer To: ▾
Download ZIP	65	3,149	161,485	Transfer To: ▾
Download ZIP	60	2,759	75,869	Transfer To: ▾
Download ZIP	55	2,461	38,497	Transfer To: ▾
Download ZIP	50	2,247	26,146	Transfer To: ▾
Download ZIP	45	2,113	22,903	Transfer To: ▾
Download ZIP	40	1,937	19,798	Transfer To: ▾

Download Network Statistics as Table

New to Cytoscape?



UniProt ID SSN for the GRE Superfamily. As described in the previous sections, the UniProt ID SSN for the GRE superfamily was generated using an alignment score threshold of 240 and a minimum length of 650 residues. The full SSN was opened with Cytoscape 3.9.1 and displayed with the yFiles Organic layout using a Mac Pro computer with 1.5TB RAM. The SSN contains 20,089 UniProt ID nodes and 47,499,276 edges.

UniRef90 Cluster SSN

The UniRef90 cluster SSN was generated for the complete UniProt entries with the **EFI-EST Family Option** by entering IPR004184 into the **Pfam and/or InterPro Families and/or Pfam clans** box (red arrow), selecting **UniRef90 cluster ID sequences** (orange arrow), checking the **Fragments** box in the **Fragment Option** to exclude fragments (blue arrow), entering the **Job name** (green arrow) and an **E-mail address** (magenta arrow), and clicking “**Submit analysis**” (black arrow).

Previous Jobs Sequence BLAST **Families** FASTA Accession IDs SSN Utilities

Generate a SSN for a protein family.

The members of the input Pfam families, InterPro families, and/or Pfam clans are selected from the UniProt, UniRef90, or UniRef50 database.

Pfam and/or InterPro Families and/or Pfam clans:

IPR004184 

Use **UniRef90** cluster ID sequences of UniProt IDs (UniProt is default). 

Family	Family Name	Full Size	UniRef90 Size	UniRef50 Size
IPR004184	PFL_dom	25,513	8,545	1,869
Total:		25,513	8,545	1,869
Total Computed:		8,545		

The input format is a single family or comma/space separated list of families. Families should be specified as PFxxxxx (five digits), IPRxxxxxx (six digits) or CLxxxx (four digits) for Pfam clans.

UniRef90 clusters contain UniProt IDs that share $\geq 90\%$ sequence identity and have 80% overlap with the longest sequence in the cluster ("seed sequence"); as a result, the UniProt IDs in the cluster usually are functionally homogeneous, i.e., orthologues.
 UniRef50 clusters contain UniProt IDs that share $\geq 50\%$ sequence identity and have 80% overlap with the seed sequence; as a result, the UniProt IDs in the cluster often are functionally heterogeneous, e.g., paralogues.

The sequences from the UniRef90 and UniRef50 databases are the UniRef90 and UniRef50 clusters for which the cluster ID ("representative sequence") matches the specified families. The UniProt members in these UniRef90 and UniRef50 clusters that do not match the specified families are removed from the cluster.

Fragment Option

UniProt designates a Sequence Status for each member: Complete if the encoding DNA sequence has both start and stop codons; Fragment if the start and/or stop codon is missing. Approximately 10% of the entries in UniProt are fragments.

Fragments: Check to exclude defined fragments in the results. (default: off) 

For the UniRef90 and UniRef50 databases, clusters are excluded if the cluster ID ("representative sequence") is a fragment.

UniProt IDs in UniRef90 and UniRef50 clusters with complete cluster IDs are removed from the clusters if they are fragments.

Filter by Taxonomy

Protein Family Size Options

Family Domain Boundary Option

SSN Edge Calculation Option

Job name:  (required)

E-mail address: 

You will be notified by e-mail when your submission has been processed.



The SSN was finalized on the **SSN Finalization** tab of the **DATASET COMPLETED** page using 240 as the **Alignment Score Threshold** that separates the SwissProt-curated functions into different clusters (orange arrow) and 650 residues as the **Minimum** in the **Sequence Length Restriction** to remove truncated sequences (cyan arrow) [1], entering the **Network (SSN) name** (brown arrow), and clicking “**Create SSN**” (black arrow).

DATASET COMPLETED

Submission Name: **IP91_IPR004184_UniRef90_NoFragments**

A minimum sequence similarity threshold that specifies the sequence pairs connected by edges is needed to generate the SSN. This threshold also determines the segregation of proteins into clusters. The threshold is applied to the edges in the SSN using the alignment score, an edge node attribute that is a measure of the similarity between sequence pairs.

Dataset Summary

Taxonomy Sunburst

Dataset Analysis

SSN Finalization

This tab is used to specify the minimum "Alignment Score Threshold" (that is a measure of the minimum sequence similarity threshold) for drawing the edges that connect the proteins (nodes) in the SSN.

Alignment Score Threshold: ?

This value corresponds to the lower limit for which an edge will be present in the SSN. The alignment score is similar in magnitude to the negative base-10 logarithm of a BLAST e-value.

▾ Sequence Length Restriction Options

Allows restriction of sequences in the generated SSN based on their length. ?

Minimum: (default: 0)

Maximum: (default: 50000)

▸ Filter by Taxonomy

▸ Neighborhood Connectivity

▸ Fragment Option

Network name: This name will be displayed in Cytoscape.

You will be notified by e-mail when the SSN is ready for download.

Create SSN

The **Network Files** tab of the **DOWNLOAD NETWORK FILES** page provided the xgmml file for the **Full (SSN) Network** (red arrow; all UniProt nodes and edges) as well as the xgmml files for **Representative Node Networks** that conflate the UniProt nodes based on percent identity (blue arrow). These files are available for download and/or transfer to the **Color SSNs** utility, **Cluster Analysis** utility, **Neighborhood Connectivity** utility, and/or EFI-GNT using the “**Transfer To**” menus (green arrows).

DOWNLOAD NETWORK FILES

Submission Name: IP91_IPR004184_UniRef90_NoFragments
 Network Name: IP91_IPR004184_UniRef90_NoFragments_Minlen650_AS240

SSN Overview Network Files

Please cite your use of the EFI tools:

Rémi Zallot, Nils Oberg, and John A. Gerlt. **The EFI Web Resource for Genomic Enzymology Tools: Leveraging Protein, Genome, and Metagenome Databases to Discover Novel Enzymes and Metabolic Pathways.** *Biochemistry* 2019 58 (41), 4169-4182. <https://doi.org/10.1021/acs.biochem.9b00735>

The panels below provide files for full and representative node SSNs for download with the indicated numbers of nodes and edges. As an approximate guide, SSNs with ~2M edges can be opened with 16 GB RAM, ~5M edges can be opened with 32 GB RAM, ~10M edges can be opened with 64 GB RAM, ~20M edges can be opened with 128 GB RAM, ~40M edges can be opened with 256 GB RAM, and ~120M edges can be opened with 768 GB RAM.

Files may be transferred to the Genome Neighborhood Tool (GNT), the Color SSN utility, the Cluster Analysis utility, the Neighborhood Connectivity utility.

Full Network ← ↓

Each node in the network represents a single protein sequence.

	# Nodes	# Edges	
Download ZIP	5,801	2,133,174	Transfer To: ▼

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

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

	% ID	# Nodes	# Edges	
Download ZIP	100	5,801	2,133,174	Transfer To: ▼
Download ZIP	95	5,772	2,098,397	Transfer To: ▼
Download ZIP	90	5,681	1,985,278	Transfer To: ▼
Download ZIP	85	4,964	1,284,066	Transfer To: ▼
Download ZIP	80	4,264	776,890	Transfer To: ▼
Download ZIP	75	3,708	444,354	Transfer To: ▼
Download ZIP	70	3,184	228,973	Transfer To: ▼
Download ZIP	65	2,762	110,675	Transfer To: ▼
Download ZIP	60	2,400	46,366	Transfer To: ▼
Download ZIP	55	2,162	25,507	Transfer To: ▼
Download ZIP	50	1,983	17,835	Transfer To: ▼
Download ZIP	45	1,879	15,914	Transfer To: ▼
Download ZIP	40	1,728	14,513	Transfer To: ▼

Download Network Statistics as Table

New to Cytoscape?

The xgmml file for the full UniRef90 cluster SSN was transferred from the **DOWNLOAD NETWORK FILES** page to the **Color SSN** utility of the **SSN Utilities** tab by clicking the **“Transfer To”** button and selecting the **Color SSN** option. This utility assigns unique numbers to each cluster (**Sequence Count Cluster Number** node attribute based on decreasing number of UniProt IDs and **Cluster Count Node Number** node attribute based on decreasing number of nodes in each cluster) and colors to the nodes in each cluster. The job was submitted by clicking **“Submit Analysis”** (black arrow).

Previous Jobs | Sequence BLAST | Families | FASTA | Accession IDs | **SSN Utilities**

Color SSNs | Cluster Analysis | Neighborhood Connectivity | Convergence Ratio

Clusters in the submitted SSN are identified, numbered and colored. Summary tables, sets of IDs and sequences per cluster are provided.

The clusters are numbered and colored using two conventions: 1) **Sequence Count Cluster Number** assigned in order of decreasing number of UniProt IDs in the cluster; 2) **Node Count Cluster Number** assigned in order of decreasing number of nodes in the cluster.

SSN File: ?

78201_IP91_IPR004184_UniRef90_NoFragments_Minlen650_AS240_full_ssn.xgmml.zip

A Cytoscape-edited SNN can serve as input. The accepted format is XGMML (or compressed XGMML as zip).

E-mail address:

You will be notified by e-mail when your submission has been processed.

Submit Analysis ←

The xgmml file for the Color SSN that was generated was downloaded from the **Data File Download** tab of the **DOWNLOAD COLORED SSN FILES** page (red arrow), opened with Cytoscape 3.9.1, and displayed with the yFiles Organic layout to obtain the Color SNN shown in **Figure 1B**.

DOWNLOAD COLORED SSN FILES

Uploaded Filename:
90691_IP91_IPR004184_UniRef90_NoFragments_Minlen650_AS240_full_ssn.xgmml

Six node attributes were added to the input SSN: **Cluster Sequence Count**, **Sequence Count Cluster Number**, **Cluster Node Count**, **Node Count Cluster Number**, **node.fillColor** (according to Cluster Sequence Count, hexadecimal), and **Node Count Fill Color** (according to Cluster Node Count, hexadecimal).

The **Data File Download** tab provides the Color SSN with the nodes colored according to **Cluster Sequence Count** (**node.fillColor**).

To change the node colors in Cytoscape to **Node Count Fill Color**: 1) select all nodes; 2) on the Style Panel, click on the "?" in the Fill Color Property; 3) select "Remove Bypass"; 4) deselect the nodes (now default node color); and 5) open the Fill Color Property and select "**Node Count Fill Color**" as the Column and "Passthrough Mapping" as the Mapping Type. The nodes will be colored with the **Node Count Fill Color**.

The **Data File Download** tab also provides files for 1) the UniProt ID-Color-Cluster Number mapping table, 2) ID Lists and FASTA Files for each cluster, 3) cluster sizes, and 4) SwissProt annotations for clusters and singletons.

Submission Summary
Data File Download

Please cite your use of the EFI tools:

Rémi Zallot, Nils Oberg, and John A. Gerlt, **The EFI Web Resource for Genomic Enzymology Tools: Leveraging Protein, Genome, and Metagenome Databases to Discover Novel Enzymes and Metabolic Pathways**. *Biochemistry* 2019 58 (41), 4169-4182. <https://doi.org/10.1021/acs.biochem.9b00735>

Colored SSN

Each cluster in the submitted SSN has been identified and assigned a unique number and color.

Download ZIP (48 MB)

Supplementary Files

Mapping Tables

Download	UniProt ID-Color-Cluster number mapping table	
----------	---	--

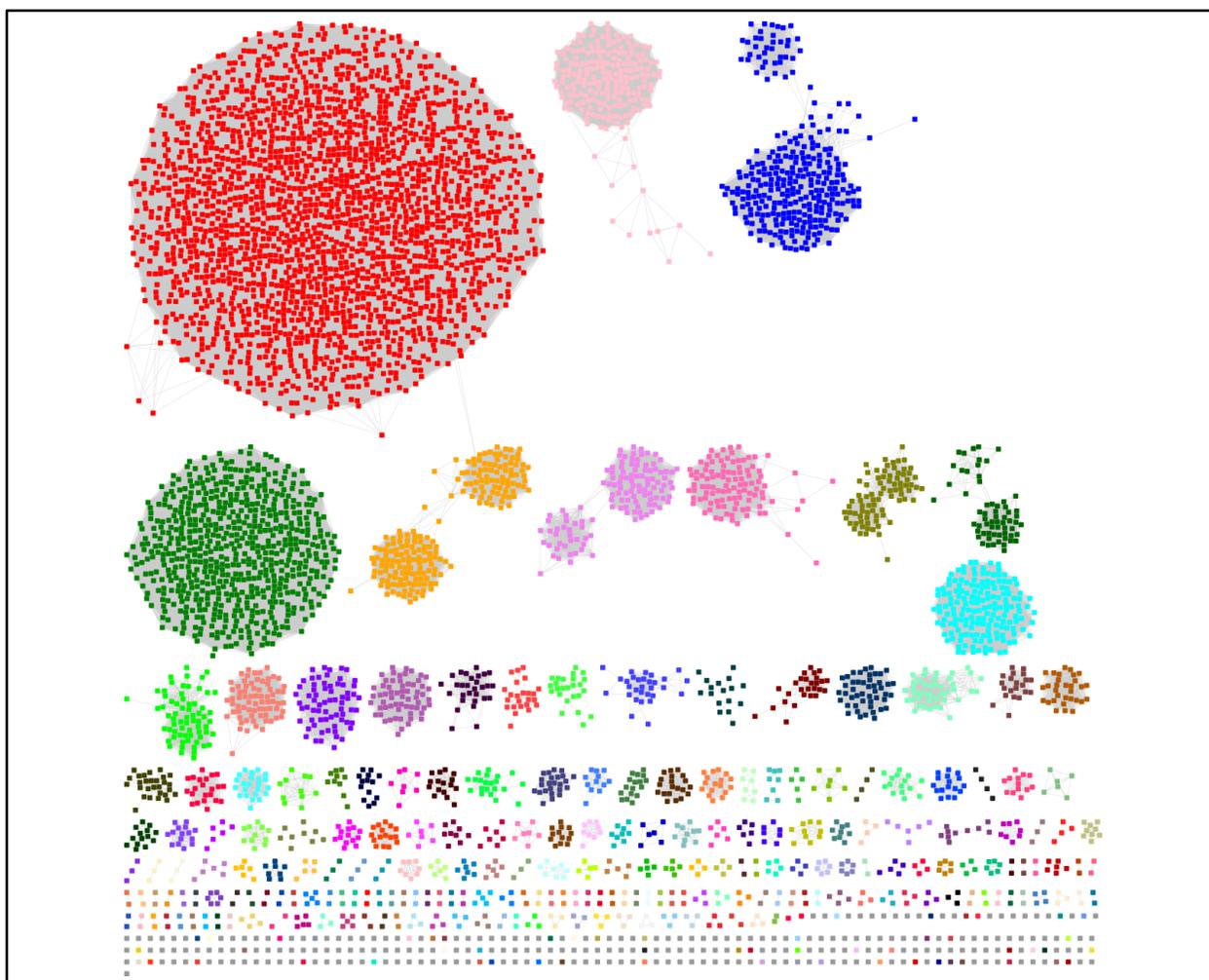
ID Lists and FASTA Files per Cluster

Download	UniProt ID lists per cluster
Download	UniRef90 ID lists per cluster
Download	FASTA files per UniProt cluster
Download	FASTA files per UniRef90 cluster

Miscellaneous Files

Download	Cluster sizes
Download	Cluster-based convergence ratio for UniProt IDs
Download	SwissProt annotations by cluster

Run CGFP on Colored SSN

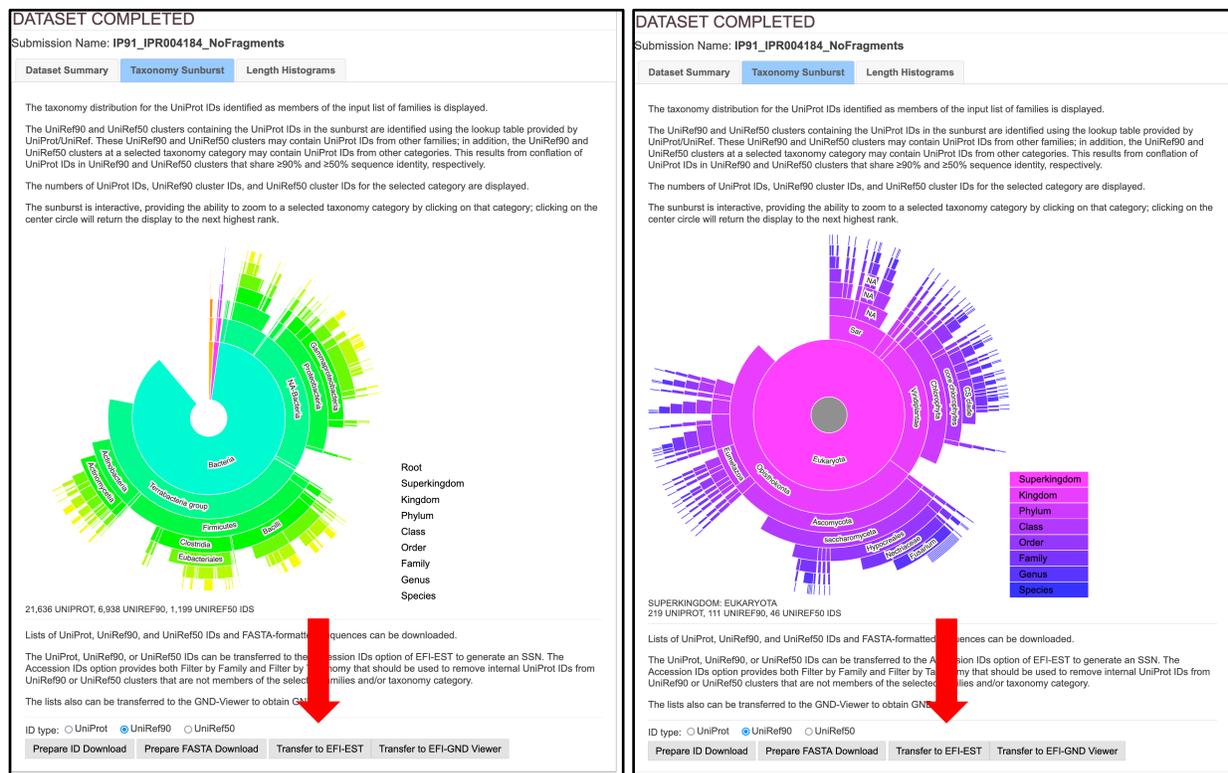


UniRef90 Cluster SSN for the GRE Superfamily. As described in the previous sections, the UniRef90 cluster SSN for the GRE superfamily was generated using an alignment score threshold of 240 and a minimum length of 650 residues. The nodes were colored using the **Color SSNs** utility. The SSN contains 5,801 UniRef90 cluster nodes and 2,133,174 edges.

The **DOWNLOAD COLORED SSN FILES** page provides other files for download, including the **UniProt ID-Color-Cluster number mapping table** (blue arrow) that can be used by the BridgeDb application in Cytoscape to color the nodes and assign cluster numbers in other SSNs that contain the same (or a subset of the same) UniProt/UniRef90/UniRef50 IDs; this file was used to color the SSNs for the taxonomy-filtered UniRef90 cluster SSNs described in the following sections.

Taxonomy Category-Specific UniRef90 SSNs: Taxonomy Tool Families Option, with transfer of UniRef90 cluster IDs to the EFI-EST Accession IDs Option

The **Taxonomy Sunburst** for complete sequences was used with the **Transfer to EFI-EST** feature (red arrow) to generate taxonomy category-specific UniRef90 SSNs. For superkingdom Bacteria; superkingdom Bacteria, phylum Actinobacteria; superkingdom Bacteria, phylum Bacteroidetes; superkingdom Bacteria, phylum Firmicutes; superkingdom Bacteria, Phylum proteobacteria; and superkingdom Archaea, the taxonomy categories were selected by clicking on the wedge (left panel). For **Preselected conditions** Fungi (four phyla within superkingdom Eukaryota), the Eukaryota taxonomy category was selected (right panel).



For the single taxonomy categories (left panel), in the **EFI-EST Accession IDs Option** pages that opened, the **Fragment Option** was used to exclude fragments (blue arrow), **Filter by Family** was used to select IPR004184 (green arrow), and **Filter by Taxonomy** was used to select the six single taxonomy categories (magenta arrow). For Fungi (right panel), **Fungi** was selected from the **Preselected conditions** menu (magenta arrow). As described in the text, **Filter by Family** and **Filter by Taxonomy** are used to ensure that the UniRef90 cluster IDs and internal UniProt IDs match the desired taxonomy category and family. The **Job name** (orange arrow) and an **E-mail address** (cyan arrow) were entered, and the job was started by clicking “**Create SSN**” (black arrow).

Previous Jobs Sequence BLAST Families FASTA **Accession IDs** SSN Utilities

Generate a SSN from a list of UniProt, UniRef, NCBI, or Genbank IDs.
An all-by-all BLAST (?) is performed to obtain the similarities between sequence pairs to calculate edge values to generate the SSN.

Use UniProt IDs Use UniRef50 or UniRef90 Cluster IDs

Input a list of UniRef50 or UniRef90 cluster accession IDs, or upload a text file.

Accession IDs:

Accession ID File:

Input accession IDs are: UniRef90 cluster IDs

Fragment Option

UniProt designates a Sequence Status for each member: Complete if the encoding DNA sequence has both start and stop codons; Fragment if the start and/or stop codon is missing. Approximately 10% of the entries in UniProt are fragments.

Fragments: Check to exclude UniProt-defined fragments. (default: off)

For the UniRef90 and UniRef50 databases, clusters are excluded if the cluster ID ("representative sequence") is a fragment.

UniProt IDs in UniRef90 and UniRef50 clusters with complete cluster IDs are removed from the clusters if they are fragments.

Filter by Family

The input list of UniRef90 or UniRef50 cluster IDs should (must) be filtered with the same list of Pfam families, InterPro families, and/or Pfam clans used to generate the IDs, if:

The input list of UniRef90 or UniRef50 IDs is obtained from 1) the Color SSN or Cluster Analysis utility for a Families option (Option B) EF-EST SSN, 2) the Families option of the Taxonomy Tool, or 3) the Accession IDs option of the Taxonomy Tool.

Input a list of Pfam families, InterPro families, and/or Pfam clans to restrict the UniProt and/or UniRef IDs in the SSN to these families.

Family(s):

The input format is a single family or comma/space separated list of families. Families should be specified as PFxxxxx (five digits), IPRxxxxxx (six digits) or CLxxxx (four digits) for Pfam clans.

For input lists of UniRef90 and UniRef50 clusters, the cluster ID (representative sequence) is used to identify those that match the list of families and are included in the SSN. The UniProt members in these clusters that do not match the input families are removed from the cluster and are not included in the SSN node attributes.

Filter by Taxonomy

The input list of UniRef90 or UniRef50 cluster IDs should (must) be filtered with the same taxonomy categories used to generate the IDs, if:

The input list of UniRef90 or UniRef50 IDs is obtained from 1) the Color SSN or Cluster Analysis utility for a Families option (Option B) EF-EST SSN, 2) the Families option of the Taxonomy Tool, or 3) the Accession IDs option of the Taxonomy Tool.

From preselected conditions, the user can select "Bacteria, Archaea, Fungi", "Eukaryota, no Fungi", "Fungi", "Viruses", "Bacteria", "Eukaryota", or "Archaea" to restrict the UniProt IDs in the subnursery to these taxonomy groups.

"Bacteria, Archaea, Fungi", "Bacteria", "Archaea", and "Fungi" select organisms that may provide genome context (gene clusters/operons) useful for inferring functions.

The UniProt IDs also can be restricted to taxonomy categories within the Superkingdom, Kingdom, Phylum, Class, Order, Family, Genus, and Species ranks. Multiple conditions are combined to be a union of each other.

Preselected conditions:

Superkingdom:

Reset

Protein Family Addition Options

Family Domain Boundary Options

SSN Edge Calculation Option

Job name: (required)

E-mail address:

You will be notified by e-mail when your submission has been processed.

Submit Analysis

Previous Jobs Sequence BLAST Families FASTA **Accession IDs** SSN Utilities

Generate a SSN from a list of UniProt, UniRef, NCBI, or Genbank IDs.
An all-by-all BLAST (?) is performed to obtain the similarities between sequence pairs to calculate edge values to generate the SSN.

Use UniProt IDs Use UniRef50 or UniRef90 Cluster IDs

Input a list of UniRef50 or UniRef90 cluster accession IDs, or upload a text file.

Accession IDs:

Accession ID File:

Input accession IDs are: UniRef90 cluster IDs

Fragment Option

UniProt designates a Sequence Status for each member: Complete if the encoding DNA sequence has both start and stop codons; Fragment if the start and/or stop codon is missing. Approximately 10% of the entries in UniProt are fragments.

Fragments: Check to exclude UniProt-defined fragments. (default: off)

For the UniRef90 and UniRef50 databases, clusters are excluded if the cluster ID ("representative sequence") is a fragment.

UniProt IDs in UniRef90 and UniRef50 clusters with complete cluster IDs are removed from the clusters if they are fragments.

Filter by Family

The input list of UniRef90 or UniRef50 cluster IDs should (must) be filtered with the same list of Pfam families, InterPro families, and/or Pfam clans used to generate the IDs, if:

The input list of UniRef90 or UniRef50 IDs is obtained from 1) the Color SSN or Cluster Analysis utility for a Families option (Option B) EF-EST SSN, 2) the Families option of the Taxonomy Tool, or 3) the Accession IDs option of the Taxonomy Tool.

Input a list of Pfam families, InterPro families, and/or Pfam clans to restrict the UniProt and/or UniRef IDs in the SSN to these families.

Family(s):

The input format is a single family or comma/space separated list of families. Families should be specified as PFxxxxx (five digits), IPRxxxxxx (six digits) or CLxxxx (four digits) for Pfam clans.

For input lists of UniRef90 and UniRef50 clusters, the cluster ID (representative sequence) is used to identify those that match the list of families and are included in the SSN. The UniProt members in these clusters that do not match the input families are removed from the cluster and are not included in the SSN node attributes.

Filter by Taxonomy

The input list of UniRef90 or UniRef50 cluster IDs should (must) be filtered with the same taxonomy categories used to generate the IDs, if:

The input list of UniRef90 or UniRef50 IDs is obtained from 1) the Color SSN or Cluster Analysis utility for a Families option (Option B) EF-EST SSN, 2) the Families option of the Taxonomy Tool, or 3) the Accession IDs option of the Taxonomy Tool.

From preselected conditions, the user can select "Bacteria, Archaea, Fungi", "Eukaryota, no Fungi", "Fungi", "Viruses", "Bacteria", "Eukaryota", or "Archaea" to restrict the UniProt IDs in the subnursery to these taxonomy groups.

"Bacteria, Archaea, Fungi", "Bacteria", "Archaea", and "Fungi" select organisms that may provide genome context (gene clusters/operons) useful for inferring functions.

The UniProt IDs also can be restricted to taxonomy categories within the Superkingdom, Kingdom, Phylum, Class, Order, Family, Genus, and Species ranks. Multiple conditions are combined to be a union of each other.

Preselected conditions:

Phylum:

Phylum:

Phylum:

Phylum:

Reset

Protein Family Addition Options

Family Domain Boundary Options

SSN Edge Calculation Option

Job name: (required)

E-mail address:

You will be notified by e-mail when your submission has been processed.

Submit Analysis

The SSNs were finalized on the **SSN Finalization** tabs of the **DATASET COMPLETED** pages using 240 as the **Alignment Score Threshold** that separates the SwissProt-curated functions into different clusters (orange arrow) and 650 residues as the **Minimum in the Sequence Length Restriction** to remove truncated sequences (cyan arrow) [1], entering the **Network (SSN) name** (brown arrow), and clicking **"Create SSN"** (black arrow).

DATASET COMPLETED

Submission Name: **IP91_IPR004184_NoFragments_Bacteria_UniRef90_NoFragments_IPR004184_Bacteria**

A minimum sequence similarity threshold that specifies the sequence pairs connected by edges is needed to generate the SSN. This threshold also determines the segregation of proteins into clusters. The threshold is applied to the edges in the SSN using the alignment score, an edge node attribute that is a measure of the similarity between sequence pairs.

This tab is used to specify the minimum "Alignment Score Threshold" (that is a measure of the minimum sequence similarity threshold) for drawing the edges that connect the proteins (nodes) in the SSN.

Alignment Score Threshold: 

This value corresponds to the lower limit for which an edge will be present in the SSN. The alignment score is similar in magnitude to the negative base-10 logarithm of a BLAST e-value.

Sequence Length Restriction Options

Allows restriction of sequences in the generated SSN based on their length. 

Minimum: (default: 0) 

Maximum: (default: 50000)

Network name: This name  played in Cytoscape.

You will be notified by e-mail when the SSN is ready for download.



The **Network Files** tab of the **DOWNLOAD NETWORK FILES** pages provided the xgmml file for the **Full (SSN) Network** (red arrow; all UniProt nodes and edges) as well as the xgmml files for **Representative Node Networks** that conflate the UniProt nodes based on percent identity (blue arrow). The xgmml files for the full SSNs were download and opened with Cytoscape; the nodes were colored using the BridgeDb app and the UniProt ID-Color-Cluster number color mapping table obtained for the Color SSN in **Figure 1B**.

DOWNLOAD NETWORK FILES

Submission Name: **IP91_IPR004184_NoFragments Bacteria UniRef90_NoFragments_IPR004184_Bacteria**

Network Name:

IP91_IPR004184_NoFragments_Bacteria_UniRef90_NoFragments_IPR004184_Bacteria_Minlen650_AS240

SSN Overview

Network Files

Please cite your use of the EFI tools:

Rémi Zallot, Nils Oberg, and John A. Gerlt, **The EFI Web Resource for Genomic Enzymology Tools: Leveraging Protein, Genome, and Metagenome Databases to Discover Novel Enzymes and Metabolic Pathways**. *Biochemistry* 2019 58 (41), 4169–4182. <https://doi.org/10.1021/acs.biochem.9b00735>

The panels below provide files for full and representative node SSNs for download with the indicated numbers of nodes and edges. As an approximate guide, SSNs with ~2M edges can be opened with 16 GB RAM, ~5M edges can be opened with 32 GB RAM, ~10M edges can be opened with 64 GB RAM, ~20M edges can be opened with 128 GB RAM, ~40M edges can be opened with 256 GB RAM, and ~120M edges can be opened with 768 GB RAM.

Files may be transferred to the Genome Neighborhood Tool (GNT), the Color SSN utility, the Cluster Analysis utility, or the Neighborhood Connectivity utility.

Full Network

Each node in the network represents a single protein sequence.

	# Nodes	# Edges	
Download ZIP	5,419	2,021,943	Transfer To: ▼

Representative Node Networks

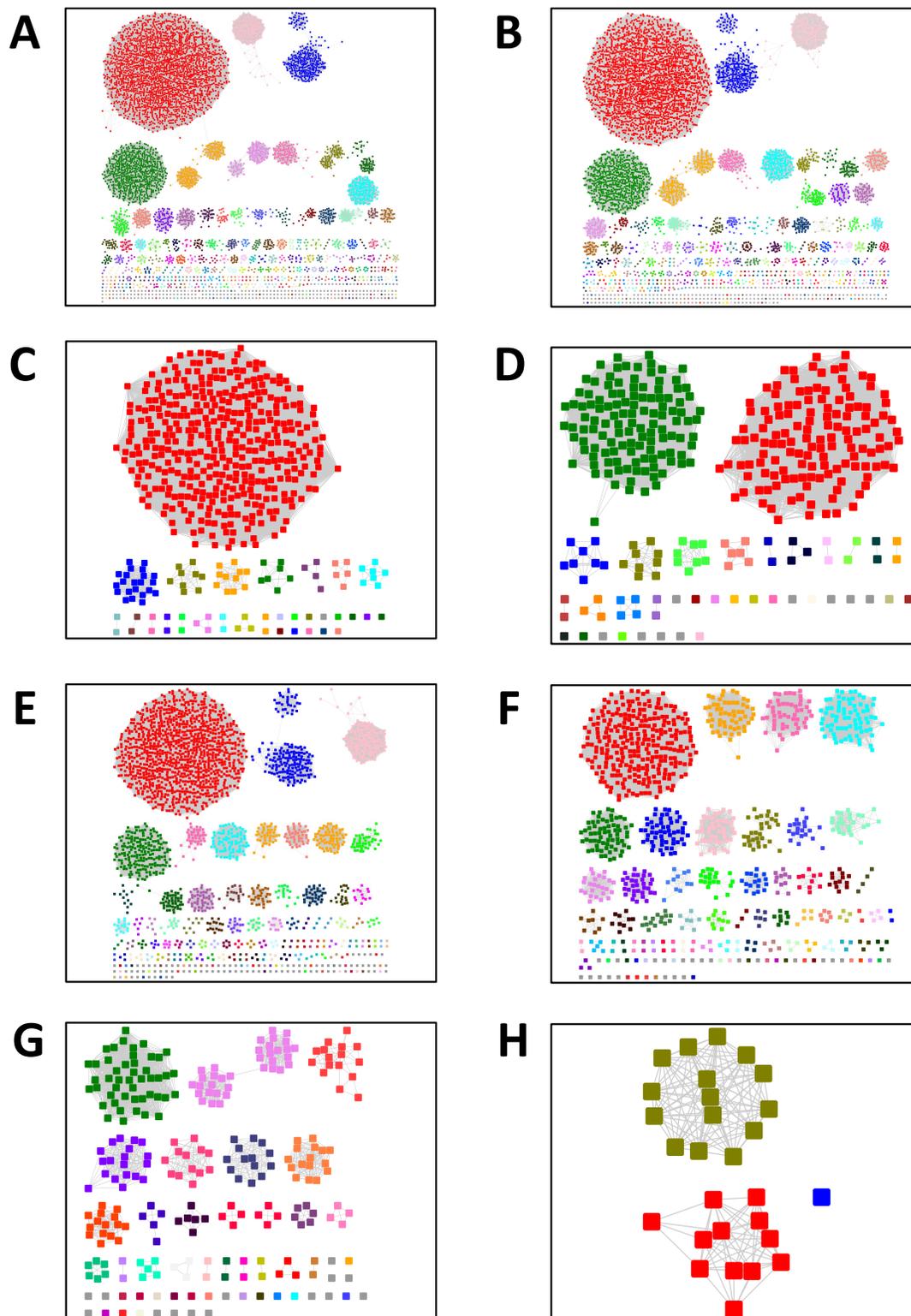
In representative node (RepNode) networks, each node in the network represents a collection of proteins grouped according to percent identity. For example, for a 75% identity RepNode network, all connected sequences that share 75% or more identity are grouped into a single node (meta node). Sequences are collapsed together to reduce the overall number of nodes, making for less complicated networks easier to load in Cytoscape.

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

	% ID	# Nodes	# Edges	
Download ZIP	100	5,419	2,021,943	Transfer To: ▼
Download ZIP	95	5,390	1,987,868	Transfer To: ▼
Download ZIP	90	5,302	1,878,667	Transfer To: ▼
Download ZIP	85	4,613	1,201,837	Transfer To: ▼
Download ZIP	80	3,934	717,327	Transfer To: ▼
Download ZIP	75	3,395	400,907	Transfer To: ▼
Download ZIP	70	2,899	202,804	Transfer To: ▼
Download ZIP	65	2,499	95,145	Transfer To: ▼
Download ZIP	60	2,159	38,269	Transfer To: ▼
Download ZIP	55	1,938	21,069	Transfer To: ▼
Download ZIP	50	1,783	14,816	Transfer To: ▼
Download ZIP	45	1,685	13,036	Transfer To: ▼
Download ZIP	40	1,551	11,871	Transfer To: ▼

Download Network Statistics as Table

[New to Cytoscape?](#)



Taxonomy Category-Filtered UniRef90 Cluster SSNs for the GRE Superfamily. The SSNs were generated using an alignment score threshold of 240 and a minimum length of 650 residues. For Panel A, the nodes were colored using the **Color SSNs** utility; for Panels B through H, the clusters/nodes were colored using the UniProt ID-Color-Cluster number color mapping table for the Color SSN in Panel A to allow the clusters/nodes to be associated with the clusters/nodes in the SSN for the entire superfamily in panel A. **Panel A**, SSN for the entire GRE superfamily; the SSN contains 5,801 nodes and 2,133,174 edges. **Panel B**, Superkingdom Bacteria; the SSN contains 5,419 nodes and 2,021,943 edges. **Panel C**, Superkingdom Bacteria, phylum Actinobacteria; the SSN contains 488 nodes and 64,199 edges. **Panel D**, Superkingdom Bacteria, phylum Bacteroidetes; the SSN contains 333 nodes and 13,658 edges. **Panel E**, Superkingdom Bacteria, phylum Firmicutes; the SSN contains 2,467 nodes and 515,667 edges. **Panel F**, Superkingdom Bacteria, phylum Proteobacteria; the SSN contains 1,048 nodes and 50,190 edges. **Panel G**, Superkingdom Archaea; the SSN contains 262 nodes and 2,074 edges. **Panel H**, Superkingdom Eukaryota, Fungi only; the SSN contains 28 nodes

Taxonomy Category-Specific UniRef90 SSNs: EFI-EST Families Option, Filter by Taxonomy in the Analysis Step

The **SSN Finalization** tab of the **DATASET COMPLETED** page for the UniRef90 cluster SSN for the complete entries was used to generate the taxonomy category-filtered SSNs described in the previous section. The SSNs were finalized using 240 as the **Alignment Score Threshold** (red arrow) and 650 residues as the **Minimum** in the **Sequence Length Restriction** (blue arrow). For the single taxonomy categories (left panel), **Filter by Taxonomy** was used to select the single taxonomy categories (green arrow). For Fungi (right panel), **Fungi** was selected from the **Preselected conditions** menu (green arrow).

DATASET COMPLETED

Submission Name: IP91_IPR004184_UniRef90_NoFragments

A minimum sequence similarity threshold that specifies the sequence pairs connected by edges is needed to generate the SSN. This threshold also determines the segregation of proteins into clusters. The threshold is applied to the edges in the SSN using the alignment score, an edge node attribute that is a measure of the similarity between sequence pairs.

Dataset Summary | Taxonomy Sunburst | Dataset Analysis | **SSN Finalization** | SSNs Created From this Dataset

This tab is used to specify the minimum "Alignment Score Threshold" (that is a measure of the minimum sequence similarity threshold) for drawing the edges that connect the proteins (nodes) in the SSN.

Alignment Score Threshold: 

This value corresponds to the lower limit for which an edge will be present in the SSN. The alignment score is similar in magnitude to the negative base-10 logarithm of a BLAST e-value.

Sequence Length Restriction Options

Allows restriction of sequences in the generated SSN based on their length. 

Minimum: (default: 0)

Maximum: (default: 50000)

Filter by Taxonomy

From preselected conditions, the user can select "Bacteria, Archaea, Fungi", "Eukaryota, no Fungi", "Fungi", "Viruses", "Bacteria", "Eukaryota", or "Archaea" to restrict the SSN nodes to these taxonomy groups.

"Bacteria, Archaea, Fungi", "Bacteria", "Archaea", and "Fungi" select organisms that may provide genome context (gene clusters/operons) useful for inferring functions.

The SSN nodes also can be restricted to taxonomy categories within the Superkingdom, Kingdom, Phylum, Class, Order, Family, Genus, and Species ranks. Multiple conditions are combined to be a union of each other.

The SSN nodes from the UniRef90 and UniRef50 databases are the UniRef90 and UniRef50 clusters for which the cluster ID ("representative sequence") matches the specified taxonomy categories. The UniProt members in these nodes that do not match the specified taxonomy categories are removed from the nodes.

Preselected conditions: 

Superkingdom:

Reset

Neighborhood Connectivity

Fragment Option

Network name: This name will be displayed in Cytoscape.

You will be notified by e-mail when the SSN is ready for download.

Create SSN

DATASET COMPLETED

Submission Name: IP91_IPR004184_UniRef90_NoFragments

A minimum sequence similarity threshold that specifies the sequence pairs connected by edges is needed to generate the SSN. This threshold also determines the segregation of proteins into clusters. The threshold is applied to the edges in the SSN using the alignment score, an edge node attribute that is a measure of the similarity between sequence pairs.

Dataset Summary | Taxonomy Sunburst | Dataset Analysis | **SSN Finalization** | SSNs Created From this Dataset

This tab is used to specify the minimum "Alignment Score Threshold" (that is a measure of the minimum sequence similarity threshold) for drawing the edges that connect the proteins (nodes) in the SSN.

Alignment Score Threshold: 

This value corresponds to the lower limit for which an edge will be present in the SSN. The alignment score is similar in magnitude to the negative base-10 logarithm of a BLAST e-value.

Sequence Length Restriction Options

Allows restriction of sequences in the generated SSN based on their length. 

Minimum: (default: 0)

Maximum: (default: 50000)

Filter by Taxonomy

From preselected conditions, the user can select "Bacteria, Archaea, Fungi", "Eukaryota, no Fungi", "Fungi", "Viruses", "Bacteria", "Eukaryota", or "Archaea" to restrict the SSN nodes to these taxonomy groups.

"Bacteria, Archaea, Fungi", "Bacteria", "Archaea", and "Fungi" select organisms that may provide genome context (gene clusters/operons) useful for inferring functions.

The SSN nodes also can be restricted to taxonomy categories within the Superkingdom, Kingdom, Phylum, Class, Order, Family, Genus, and Species ranks. Multiple conditions are combined to be a union of each other.

The SSN nodes from the UniRef90 and UniRef50 databases are the UniRef90 and UniRef50 clusters for which the cluster ID ("representative sequence") matches the specified taxonomy categories. The UniProt members in these nodes that do not match the specified taxonomy categories are removed from the nodes.

Preselected conditions: 

Phylum:

Phylum:

Phylum:

Phylum:

Reset

Neighborhood Connectivity

Fragment Option

Network name: This name will be displayed in Cytoscape.

You will be notified by e-mail when the SSN is ready for download.

Create SSN

The **Network Files** tab of the **DOWNLOAD NETWORK FILES** pages provided the xgmml file for the **Full (SSN) Network** (red arrow; all UniProt nodes and edges) as well as the xgmml files for **Representative Node Networks** that conflate the UniProt nodes based on percent identity (blue arrow). The xgmml files for the full SSNs were downloaded and opened with Cytoscape; the nodes were colored using the BridgeDb app and the UniProt ID-Color-Cluster number color mapping table obtained for the Color SSN in **Figure 1B**.

DOWNLOAD NETWORK FILES

Submission Name: IP91_IPR004184_UniRef90_NoFragments
 Network Name: IP91_IPR004184_UniRef90_NoFragments_Bacteria_Minlen650_AS240

SSN Overview **Network Files**

Please cite your use of the EFI tools:

Rémi Zallot, Nils Oberg, and John A. Gerit, **The EFI Web Resource for Genomic Enzymology Tools: Leveraging Protein, Genome, and Metagenome Databases to Discover Novel Enzymes and Metabolic Pathways**. *Biochemistry* 2019 58 (41), 4169-4182. <https://doi.org/10.1021/acs.biochem.9b00735>

The panels below provide files for full and representative node SSNs for download with the indicated numbers of nodes and edges. As an approximate guide, SSNs with ~2M edges can be opened with 16 GB RAM, ~5M edges can be opened with 32 GB RAM, ~10M edges can be opened with 64 GB RAM, ~20M edges can be opened with 128 GB RAM, ~40M edges can be opened with 256 GB RAM, and ~120M edges can be opened with 768 GB RAM.

Files may be transferred to the Genome Neighborhood Tool (GNT), the Color SSN utility, the Cluster Analysis utility, or the Neighborhood Connectivity utility.

Full Network ← ?

Each node in the network represents a single protein sequence.

	# Nodes	# Edges	
Download ZIP	5,419	2,021,943	Transfer To: ▾

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

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

	% ID	# Nodes	# Edges	
Download ZIP	100	5,419	2,021,943	Transfer To: ▾
Download ZIP	95	5,390	1,987,868	Transfer To: ▾
Download ZIP	90	5,302	1,878,667	Transfer To: ▾
Download ZIP	85	4,613	1,201,837	Transfer To: ▾
Download ZIP	80	3,934	717,327	Transfer To: ▾
Download ZIP	75	3,395	400,907	Transfer To: ▾
Download ZIP	70	2,899	202,804	Transfer To: ▾
Download ZIP	65	2,499	95,145	Transfer To: ▾
Download ZIP	60	2,159	38,269	Transfer To: ▾
Download ZIP	55	1,938	21,069	Transfer To: ▾
Download ZIP	50	1,783	14,816	Transfer To: ▾
Download ZIP	45	1,685	13,036	Transfer To: ▾
Download ZIP	40	1,551	11,871	Transfer To: ▾

Download Network Statistics as Table

New to Cytoscape?

Taxonomy Category-Specific UniRef90 SSNs: EFI-EST Families Option, Filter by Taxonomy in the Generate Step

The same taxonomy category-filtered UniRef90 SSNs were generated in separate jobs using the **EFI-EST Family Option** by specifying IPR004184 as the input family (red arrow), selecting UniRef90 cluster IDs (blue arrow), selecting **Fragment Option** to exclude fragments (green arrow), and selecting the taxonomy categories (magenta arrow; single categories in the left panel; Fungi in the right panel). The **Job name** (orange arrow) and an **E-mail address** (cyan arrow) were entered; the job was started by clicking “**Submit analysis**” (black arrow).

Previous Jobs Sequence BLAST Families FASTA Accession IDs SSN Utilities

Generate a SSN for a protein family.

The members of the input Pfam families, InterPro families, and/or Pfam clans are selected from the UniProt, UniRef90, or UniRef50 database.

Pfam and/or InterPro Families and/or Pfam clans:

IPR004184

Use UniRef90 cluster IDs in sequences instead of UniProt IDs (UniProt is default).

Family	Size	UniRef90 Size	UniRef50 Size
IPR004184 PFL_dom	25,513	8,545	1,869
Total:	25,513	8,545	1,869
Total Computed:	8,545		

The input format is a single family or comma/space separated list of families. Families should be specified as PFxxxxx (five digits), IPRxxxxxx (six digits) or CLxxxx (four digits) for Pfam clans.

UniRef90 clusters contain UniProt IDs that share >90% sequence identity and have 80% overlap with the longest sequence in the cluster ("seed sequence"), as a result, the UniProt IDs in the cluster usually are functionally homogeneous, i.e., orthologues. UniRef50 clusters contain UniProt IDs that share >50% sequence identity and have 80% overlap with the seed sequence; as a result, the UniProt IDs in the cluster often are functionally heterogeneous, e.g., paralogues.

The sequences from the UniRef90 and UniRef50 databases are the UniRef90 and UniRef50 clusters for which the cluster ID ("representative sequence") matches the specified families. The UniProt members in these UniRef90 and UniRef50 clusters that do not match the specified families are removed from the cluster.

Fragment Option

UniProt designates a Sequence Status for each member. Complete if the encoding DNA sequence has both start and stop codons; Fragment if the start and/or stop codon is missing. Approximately 10% of the entries in UniProt are fragments.

Fragments: Check to exclude fragments in the results. (default: off)

For the UniRef90 and UniRef50 databases, clusters are excluded if the cluster ID ("representative sequence") is a fragment.

UniProt IDs in UniRef90 and UniRef50 clusters with complete cluster IDs are removed from the clusters if they are fragments.

Filter by Taxonomy

From preselected conditions, the user can select "Bacteria, Archaea, Fungi", "Eukaryota, no Fungi", "Fungi", "Viruses", "Bacteria", "Eukaryota", or "Archaea" to restrict the retrieved sequences to these taxonomy groups.

"Bacteria, Archaea, Fungi", "Bacteria", "Archaea", and "Fungi" select organisms that may provide genome context (gene clusters/operons) useful for inferring functions.

The retrieved sequences also can be restricted to taxonomy categories within the Superkingdom, Kingdom, Phylum, Class, Order, Family, Genus, and Species ranks. Multiple conditions are combined to be a union of each other.

The sequences from the UniRef90 and UniRef50 databases are the UniRef90 and UniRef50 clusters for which the cluster ID ("representative sequence") matches the specified taxonomy categories. The UniProt members in these clusters that do not match the specified taxonomy categories are removed from the cluster.

Preselected conditions: Bacteria

Superkingdom: Bacteria

Reset

Protein Family Size Options

Family Domain Boundary Option

SSN Edge Calculation Option

Job name: *IP91_IPR004184_UniRef90_NoFragments_Bacteria* (required)

E-mail address:

You will be notified by e-mail when your submission has been processed.

Submit Analysis

Previous Jobs Sequence BLAST Families FASTA Accession IDs SSN Utilities

Generate a SSN for a protein family.

The members of the input Pfam families, InterPro families, and/or Pfam clans are selected from the UniProt, UniRef90, or UniRef50 database.

Pfam and/or InterPro Families and/or Pfam clans:

IPR004184

Use UniRef90 cluster IDs in sequences instead of UniProt IDs (UniProt is default).

Family	Size	UniRef90 Size	UniRef50 Size
IPR004184 PFL_dom	25,513	8,545	1,869
Total:	25,513	8,545	1,869
Total Computed:	8,545		

The input format is a single family or comma/space separated list of families. Families should be specified as PFxxxxx (five digits), IPRxxxxxx (six digits) or CLxxxx (four digits) for Pfam clans.

UniRef90 clusters contain UniProt IDs that share >90% sequence identity and have 80% overlap with the longest sequence in the cluster ("seed sequence"), as a result, the UniProt IDs in the cluster usually are functionally homogeneous, i.e., orthologues. UniRef50 clusters contain UniProt IDs that share >50% sequence identity and have 80% overlap with the seed sequence; as a result, the UniProt IDs in the cluster often are functionally heterogeneous, e.g., paralogues.

The sequences from the UniRef90 and UniRef50 databases are the UniRef90 and UniRef50 clusters for which the cluster ID ("representative sequence") matches the specified families. The UniProt members in these UniRef90 and UniRef50 clusters that do not match the specified families are removed from the cluster.

Fragment Option

UniProt designates a Sequence Status for each member. Complete if the encoding DNA sequence has both start and stop codons; Fragment if the start and/or stop codon is missing. Approximately 10% of the entries in UniProt are fragments.

Fragments: Check to exclude fragments in the results. (default: off)

For the UniRef90 and UniRef50 databases, clusters are excluded if the cluster ID ("representative sequence") is a fragment.

UniProt IDs in UniRef90 and UniRef50 clusters with complete cluster IDs are removed from the clusters if they are fragments.

Filter by Taxonomy

From preselected conditions, the user can select "Bacteria, Archaea, Fungi", "Eukaryota, no Fungi", "Fungi", "Viruses", "Bacteria", "Eukaryota", or "Archaea" to restrict the retrieved sequences to these taxonomy groups.

"Bacteria, Archaea, Fungi", "Bacteria", "Archaea", and "Fungi" select organisms that may provide genome context (gene clusters/operons) useful for inferring functions.

The retrieved sequences also can be restricted to taxonomy categories within the Superkingdom, Kingdom, Phylum, Class, Order, Family, Genus, and Species ranks. Multiple conditions are combined to be a union of each other.

The sequences from the UniRef90 and UniRef50 databases are the UniRef90 and UniRef50 clusters for which the cluster ID ("representative sequence") matches the specified taxonomy categories. The UniProt members in these clusters that do not match the specified taxonomy categories are removed from the cluster.

Preselected conditions: Fungi

Phylum: Ascomycota

Phylum: Basidiomycota

Phylum: Fungi incertae sedis

Phylum: unclassified fungi

Reset

Protein Family Size Options

Family Domain Boundary Option

SSN Edge Calculation Option

Job name: *IP91_IPR004184_UniRef90_NoFragments_Fungi* (required)

E-mail address:

You will be notified by e-mail when your submission has been processed.

Submit Analysis

As described previously for the UniProt ID and UniRef90 cluster SSNs, the SSNs were finalized (**SSN Finalization** tab on the **DATASET COMPLETED** pages) using 240 as the **Alignment Score Threshold** and 650 residues as the **Minimum** in the **Sequence Length Restriction**, entering the **Network (SSN) name** (brown arrow), and clicking “**Create SSN**” (black arrow).

DATASET COMPLETED

Submission Name: **IP91_IPR004184_UniRef90_NoFragments_Bacteria**

A minimum sequence similarity threshold that specifies the sequence pairs connected by edges is needed to generate the SSN. This threshold also determines the segregation of proteins into clusters. The threshold is applied to the edges in the SSN using the alignment score, an edge node attribute that is a measure of the similarity between sequence pairs.

Dataset Summary
Taxonomy Sunburst
Dataset Analysis
SSN Finalization

This tab is used to specify the minimum "Alignment Score Threshold" (that is a measure of the minimum sequence similarity threshold) for drawing the edges that connect the proteins (nodes) in the SSN.

Alignment Score Threshold: ?

This value corresponds to the lower limit for which an edge will be present in the SSN. The alignment score is similar in magnitude to the negative base-10 logarithm of a BLAST e-value.

▾ Sequence Length Restriction Options

Allows restriction of sequences in the generated SSN based on their length. ?

Minimum: (default: 0)

Maximum: (default: 50000)

▸ Filter by Taxonomy
▸ Neighborhood Connectivity
▸ Fragment Option

Network name: This name will be used in Cytoscape.

You will be notified by e-mail when the SSN is ready for download.

Create SSN

The **Network Files** tab of the **DOWNLOAD NETWORK FILES** pages provided the xgmml file for the **Full (SSN) Network** (red arrow; all UniProt nodes and edges) as well as the xgmml files for **Representative Node Networks** that conflate the UniProt nodes based on percent identity (blue arrow). The xgmml files for the full SSNs were download and opened with Cytoscape; the nodes were colored using the BridgeDb app and the UniProt ID-Color-Cluster number color mapping table obtained for the Color SSN in **Figure 1B**.

DOWNLOAD NETWORK FILES

Submission Name: IP91_IPR004184_UniRef90_NoFragments_Bacteria
 Network Name: IP91_IPR004184_UniRef90_NoFragments_Bacteria_Minlen650_AS240

SSN Overview **Network Files**

Please cite your use of the EFI tools:

Rémi Zallot, Nils Oberg, and John A. Gerlt, **The EFI Web Resource for Genomic Enzymology Tools: Leveraging Protein, Genome, and Metagenome Databases to Discover Novel Enzymes and Metabolic Pathways.** *Biochemistry* 2019 58 (41), 4169-4182. <https://doi.org/10.1021/acs.biochem.9b00735>

The panels below provide files for full and representative node SSNs for download with the indicated numbers of nodes and edges. As an approximate guide, SSNs with ~2M edges can be opened with 16 GB RAM, ~5M edges can be opened with 32 GB RAM, ~10M edges can be opened with 64 GB RAM, ~20M edges can be opened with 128 GB RAM, ~40M edges can be opened with 256 GB RAM, and ~120M edges can be opened with 768 GB RAM.

Files may be transferred to the Genome Neighborhood Tool (GNT), the Color SSN utility, the Cluster Analysis utility, or the Neighborhood Connectivity utility.

Full Network ← ?

Each node in the network represents a single protein sequence.

	# Nodes	# Edges	
Download ZIP	5,419	2,021,943	Transfer To: ▼

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

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

	% ID	# Nodes	# Edges	
Download ZIP	100	5,419	2,021,943	Transfer To: ▼
Download ZIP	95	5,390	1,987,868	Transfer To: ▼
Download ZIP	90	5,302	1,878,667	Transfer To: ▼
Download ZIP	85	4,613	1,201,837	Transfer To: ▼
Download ZIP	80	3,934	717,327	Transfer To: ▼
Download ZIP	75	3,395	400,907	Transfer To: ▼
Download ZIP	70	2,899	202,804	Transfer To: ▼
Download ZIP	65	2,499	95,145	Transfer To: ▼
Download ZIP	60	2,159	38,269	Transfer To: ▼
Download ZIP	55	1,938	21,069	Transfer To: ▼
Download ZIP	50	1,783	14,816	Transfer To: ▼
Download ZIP	45	1,685	13,036	Transfer To: ▼
Download ZIP	40	1,551	11,871	Transfer To: ▼

Download Network Statistics as Table

New to Cytoscape?

Tutorial 2. Use of the Taxonomy Tool and Filter by Taxonomy: RS Superfamily

This tutorial provides the details for the jobs described in the **Taxonomy Tool and Filter By Taxonomy: RS Superfamily** section in the text. The web resource **Training** page (<https://efi.igb.illinois.edu/training/example.php?id=2022>) provides links to 1) the **Taxonomy Taxonomy Tool** jobs used to generate **Taxonomy Sunbursts** and 2) both the **DATASET COMPLETED** and **DOWNLOAD NETWORK FILES** pages for the EFI-EST jobs used for generating the taxonomy category-filtered SSNs.

Taxonomy Sunbursts: Taxonomy Tool Families Option

Complete Sequences and Fragments. The **Taxonomy Sunburst** for complete sequences and fragments was generated from UniProt Release 2022_04 by entering a list of 211 Pfam and InterPro families and/or domains (**Tutorial Table 1**) into the **Pfam and/or InterPro Families and/or Pfam Clans** box on the **Taxonomy Tool Families Option** page (red arrow), entering the **Job name** (orange) and an **E-mail address** (cyan arrow), and clicking “**Submit analysis**” (black arrow). The Tool provides the list of input families/domains with the numbers of UniProt, UniRef90, and UniRef50 clusters in each, so the page is long (next page). The two boxed areas are enlarged on the following page to clearly show the input parameters.

Previous jobs: [Families](#) [FASTA](#) [Accession IDs](#)

Retrieve taxonomy for families.

The UniProt IDs for family members are identified in UniProtKB with a list of Pfam families, InterPro families, and/or Pfam clans.

Pfam and/or InterPro Families:

45784 PF04055 PF06969 PF00497 PF12345 PF13106 PF16199 PF16881 PF19238 PF19288 PF19064

Family	Family Name	Full Size	UniRef50 Size	UniRef50 Size
IR000385	MoaA_NMB_PqE_Fe-S-td_CS	49,241	23,160	4,777
IR001989	Radical_actin_CS	26,936	9,907	1,836
IR002684	Biotin_synthase	27,641	9,800	1,504
IR003888	LipoY_synth	36,047	13,904	1,318
IR003739	Lys_aminomutase_Glu_NH3_mut	20,775	10,372	1,278
IR004263	rRNA_16S_M75ase_RimN_Cp	39,844	15,429	1,450
IR004538	Coproporphyrinogen_hemK	16,796	6,746	913
IR004559	HemK-like	38,255	17,990	2,765
IR005839	Methyltransferase	87,716	37,296	4,127
IR005940	Ribosome_212_MedTfase_RimO	28,608	11,857	2,029
IR005949	MuB-like_B	4,407	2,208	508
IR006457	MuB-like_C	17,077	6,216	1,089
IR006638	EloMAnMB-like_SAM	446,282	212,389	36,535
IR007079	ISAM	722,533	355,669	70,723
IR007080	MuB_synth_C	38,281	16,656	1,862
IR007072	BAT3_dom	39,882	14,853	1,222
IR007073	HemN_C	39,466	17,002	2,804
IR001101	DUF5131	7,313	4,801	1,324
IR001943	PQQ_synth_PqqE_bac	5,549	1,635	591
IR001287	hmgD	11,644	4,135	630
IR001288	Pfl1_activating	9,426	2,995	180
IR001289	Dppase_radical_activase	16,374	5,941	1,127
IR001343	MuB	33,719	14,364	976
IR0013704	UPF0313_N	12,479	4,760	404
IR0013848	Methyltransferase_N	91,463	39,689	5,468
IR0013917	rRNA_16S_M75ase-synth	8,013	4,099	552
IR001410	rRNA_16S_activator	26	20	27
IR0014631	Pyruvate_lysase_activ_prd	16,215	11,197	1,144
IR0016771	Fe-S_OuRdase_ISAM_TM048_prd	62	39	6
IR0016779	ISAM_MSM05088	1,625	751	63
IR0016883	MuB	36	34	6
IR0017000	PqqE-like	19,308	10,212	2,373
IR0017172	MuA_551-like	433	209	15
IR0017173	Deazaganine_synth	2,879	884	27
IR0017833	Hsp90act_synth_assoc_ISAM_HspH	4,342	1,470	62
IR0017834	Hsp90act_synth_assoc_ISAM_HspH	1,453	448	14
IR001939	CuR_famly	6,705	3,144	254
IR001940	CuR_famly	6,240	2,738	154
IR0020093	FD_synthase_suf2	20,381	8,463	490
IR0020612	Methyltransferase_CS	96,820	40,909	5,265
IR002431	Cyclic_DHFR_synthase_mprC	6,120	2,294	95
IR002432	MprE	5,884	2,259	97
IR002441	Lys_aminomutase-ref	1,692	800	38
IR002459	Lysine_aminomutase	3,271	1,505	12
IR002462	EglnB	5,760	2,170	180
IR002881	rRNA_16S_M75ase_Cp	258	54	2
IR002948	UPF0313	13,008	5,108	563
IR002949	ISAM_S2L_activator	197,152	100,169	21,903
IR002955	UhoAcharacterized_Spiv1	1,028	422	9
IR002987	Paq66a_mod_ISAM	116	104	37
IR0029819	Paq6-mod_ISAM_AF057	682	468	66
IR002989	ISAM_S2L_activator	219	33	1
IR003261	ISAM_Tad_assoc_C	2,039	1,206	206
IR0032822	ISAM_Tad_assoc_bac	438	178	22
IR003288	ISAM_HmbB	136	61	8
IR003289	CHP0460_ISAM	7,088	3,804	588
IR003293	ISAM_PFD0134	82	36	5
IR0032987	Sulfolipase_maturase_ISAM	26,588	13,657	5,204
IR0032988	7-CO-2-deazaGln_synth_pul_Clo	534	250	13
IR0032989	DNA_ISAM_pul	8,922	3,262	228
IR0032990	BenzoylCoA_synthase_activating	10	8	1
IR0032985	4H-e4-binding_SPSAM_dom	60,484	34,786	13,236
IR0032986	GH-AmdH1_gou_mutation	777	235	15
IR0032991	Pyriminyl_PyB	206	172	23
IR0032992	SpoM_PP_lysase	1,810	628	25
IR0032994	Paq_ISAM_mnt_TyD0	36	16	6
IR0032992	Yjw_bact	2,077	431	55
IR0032993	MHC	1,478	492	22
IR0032995	MuB1	289	128	6
IR0032999	CHP046072_B12-tdISAM	68	60	15
IR0032979	CHP04614_B12-tdISAM	212	108	15
IR0032980	CHP04613_B12-tdISAM	910	558	151
IR0032981	ISAM_act1	11,314	407	597
IR0032982	HemO1_synth_NuJ	1,137	491	8
IR0032993	TYW1_archaea	2,475	1,376	136

IR0024001	Cys-rot_pqr_ISAM_CcpM	231	130	67
IR0024007	FerA-rot_pqr_mnt_NuYD	3,483	1,764	78
IR0024016	CHP04604_ISAM	20	14	4
IR0024017	Pqr_pqr_ISAM	45	6	2
IR0024018	CHP04603_ISAM	330	159	14
IR0024021	FerA-rot_pqr_ISAM	3,733	2,075	217
IR0024023	ISAM_pqr1_Head	1,143	625	125
IR0024025	SCIFF_ISAM_maturase	2,738	1,232	79
IR0024032	ISAM_pqr2_Head	1,002	528	142
IR0024077	ISAM_pqr3_Head	24,204	8,537	476
IR002451	DUF5141	4,310	2,732	226
IR0024500	UPF0313_C	11,093	3,850	317
IR0024824	7-CO-2-deazaganine_synth-like	25,740	11,573	1,114
IR002595	LAM_C_dom	6,846	3,760	332
IR0026302	Chaperon_mnt_ISAM	106	95	26
IR0026332	HuWV	1,992	438	75
IR0026335	SAM_SPSAM_FxkB	1,394	895	82
IR0026344	SCM_ISAM_ScmE	66	62	7
IR0026346	SCM_ISAM_ScmF	59	56	6
IR0026351	ISAM_ScCys	4,383	2,794	272
IR0026357	ISAMSPASAM_pqr1_GRRM_system	342	189	72
IR0026401	COOX_matur	139	83	30
IR0026404	ISAM_mnt_3po	429	175	16
IR0026407	SAM_GD-Bacter	267	131	67
IR0026412	ISAM_CooX_rpt	171	108	34
IR0026423	ISAM_cobpase	190	138	15
IR0026426	ISAM_Pfl1Domain	18	10	6
IR0026429	MA_synthase	13	10	1
IR0026447	B12_SAM_mnt16	457	325	66
IR0026482	ISAM_mnt1_3	129	44	1
IR0027492	rRNA_16S_HspH	37,070	13,980	1,140
IR0027500	LipoY_synth_mpr	337	132	20
IR0027527	LipoY_synth_mt	205	79	10
IR0027559	B12_ISAM_aligo	443	144	8
IR0027914	HspR_B12_ISAM	427	165	5
IR0027919	QuaeE_Cp14CoC	11	9	6
IR0027983	ISAM_ACGX	154	98	8
IR0027986	ISAM_mnt1_mat	167	146	44
IR0027990	ArmaMkMam/S	10,871	7,172	845
IR0027994	ISAM_matur	456	260	126
IR0027998	Sphe_SPSAM	330	177	83
IR0028009	ISAM_QuaeE_Proteobac	2,827	501	21
IR0028074	ISAM_QuaeE_bac	3,644	1,315	28
IR0028212	ISAM_CuRact	9	11	3
IR0028226	Pqq6a_SAM_Halo	248	67	1
IR0028333	ISAM_NuJ2	399	188	5
IR0030081	Glu_2_3_NuJ_mnt	266	178	2
IR0030837	B12_ISAM_act1	41	35	4
IR0030844	ISAM_Propionactase	570	287	4
IR0030896	ISAM_AmdH_hemB	560	364	19
IR0030905	CuC_activ_ISAM	564	190	10
IR0030915	ISAM_SuB	33	10	6
IR0030931	Mnt_1po_ISAM	34	21	4
IR0030990	ISAM_PyO	190	106	69
IR0030969	B12_SAM_1po_MT	74	46	2
IR0030977	QuaeE_Cp14CoC	3,147	1,635	23
IR0030989	ISAM_PyB1	38	16	3
IR0031003	BspO1_PhpC_ISAM	97	75	29
IR0031004	ISAM_YRkA	1,610	528	16
IR0031010	ISAM_mnt_PaA	702	29	4
IR0031012	ISAM_mnt_pqrB	802	101	4
IR0031014	ISAM_EHc	39	23	1
IR0031015	Arg_2_3_mnt_muta	160	51	4
IR0031019	ISAM_mnt_Cn	11	11	6
IR0031688	Lias_N	25,765	8,364	724
IR0032432	Radical_SAM_C	20,918	9,267	1,452
IR0033971	Aviamycin_synthase	9	3	1
IR0033974	Glycerol_dehydratase_activase	33	5	3
IR0033975	TrpH-like	9	11	3
IR0033976	GHE-like	9	1	1
IR0034166	NMB_C	2,248	884	94
IR0034386	BoH-like	3	2	2
IR0034391	Cmo-like_SPSAM_containing	9,391	6,770	2,830
IR0034405	F4D0	29,713	12,005	1,325
IR0034422	HydPylB-like	5,384	3,253	647
IR0034428	ThiNacNCLHydG-like	10,820	4,489	454
IR0034436	hsohHsaH-like	11	9	1
IR0034438	4-Hydroxy_nicotinamide_activase	15	1	1
IR0034457	Organic_radical_activating	53,598	26,055	4,922
IR0034462	BenzoylCoA_synthase	21	4	2
IR0034465	Pyruvate_lysase_activase	3,838	433	27
IR0034468	Methyltransferase_Class_B	32,110	22,756	3,051
IR0034471	GGDTMA_synthase	1,393	710	47
IR0034474	Methyltransferase_Class_D	5,445	3,164	757
IR0034479	ISAM-lik	542	477	11
IR0034480	HemC_cytohcy_lysase-like	244	202	37
IR0034485	Araeobac_Cys-type_sulfite-m	882	365	75
IR0034491	Araeobac_Ser_sulfite-maturase	5,937	1,729	148
IR0034492	ISAM-lik	59	31	1

IR0034498	Bacteriochlorophyll_Col_MT	47	33	1
IR0034502	Coproporphyrinogen-III_activase	66,641	31,743	5,800
IR0034508	Spectinomycin_synthetase	8	7	5
IR0034514	TrvK-like	11	3	1
IR0034515	TrvL-like	8	4	2
IR0034519	TrvH-like	11	6	1
IR0034529	FonK-like	5	2	1
IR0034530	Hpr-P	3,569	1,910	158
IR0034581	Methylation_of_yankermysin	9	4	1
IR0034583	TrvH_rRNA_Methytransferase	3,721	1,422	128
IR0034589	SpoM_PP_lysase_Chlostridia	537	327	20
IR0034600	SpoM_PP_lysase_Bactri	1,273	302	6
IR0034687	ELP3-like	7,115	3,099	309
IR0034818	Methyltransferase_N_uI	91,072	39,494	5,362
IR0034861	ELP3	23,268	10,427	1,876
IR0040072	Methyltransferase_A	43,050	17,542	2,586
IR0040074	BspDPTAA1YW	9,191	3,999	962
IR0040081	CndI-like	1	2	1
IR0040082	QuaeE-like	4	1	1
IR0040085	MU674-like	7,375	4,888	824
IR0040086	MU683-like	22,149	11,842	1,979
IR0040087	MU692-like	1,943	1,037	231
IR0040088	MU693-like	11,805	6,446	1,460
IR0041582	Radical_SAM_N2	25,468	10,202	1,260
IR0043375	PuI_radical_SAM-like_N	4,618	2,655	513
IR0043387	Radical_SAM_N	21,334	9,047	896
IR0043388	Radical_SAM_N2	11,805	6,446	1,460
PF04055	Radical_SAM	672,681	327,815	62,860
PF06969	HemN_C	36,495	17,002	2,804
PF08493	Radical_SAM_N	12,479	4,760	404
PF12345	DUF5141	4,310	2,732	226
PF13106	SPASAM	47,262	26,993	9,906
PF16199	Radical_SAM_C	20,918	9,267	1,452
PF18881	Lias_N	25,765	8,364	724
PF19238	Radical_SAM_2	4,815	2,605	513
PF19288	CuR_C	21,334	9,047	896
PF19964	Radical_SAM_N2	11,805	6,446	1,460
Total:		4,055,668	1,898,246	336,149

The input format is a single family or comma-separated list of families. Families should be specified as PFxxxxx (six digits), PFxxxxxx (six digits) or CLxxxx (four digits) for Pfam clans.

Filter by Taxonomy can be used to remove UniProt IDs that do not match the specified taxonomy categories. The remaining UniProt IDs are used to generate the subunit.

The UniProt IDs and UniRef50 clusters that contain the UniProt IDs are retrieved from the UniProt and UniRef50 databases using the Accession IDs provided by UniProt/UniRef50. Clusters for which the cluster ID (representative sequence) matches the list of families are retained.

The numbers of UniProt IDs and both UniRef50 cluster and UniRef50 cluster IDs are displayed on the subunit; the UniProt IDs and both UniRef50 cluster and UniRef50 cluster IDs are available for download and/or transfer to the Accession ID option (Option D) of EPI-EST to generate SIDs.

If the lists of UniRef50 or UniRef50 cluster IDs are used to generate SIDs with the Accession ID option (Option D) of EPI-EST, the lists should (must) be filtered with the same list of families (Filter by Family) and any specified taxonomy categories (Filter by Taxonomy) used to generate the lists.

This filtering removes the UniRef50 and UniRef50 clusters with cluster IDs that do not match the UniProt IDs that are not members of the specified families or have the selected taxonomy categories.

Fragment Option

UniProt designates a Sequence Status for each member. Complete if the encoding DNA sequence has both start and stop codons; Fragment if the start and/or stop codon is missing. Approximately 10% of the entries in UniProt are fragments.

Fragments: Check to exclude UniProt-defined fragments in the results. (default: off)

For the UniRef50 and UniRef50 databases, clusters are excluded if the cluster ID (representative sequence) is a fragment.

UniProt IDs in UniRef50 and UniRef50 clusters with complete cluster IDs are removed from the clusters if they are fragments.

Filter by Taxonomy

This filter is applied to the UniProt IDs after they have been identified using the list of Pfam families, InterPro families, and/or Pfam clans. The remaining UniProt IDs are used to generate the subunit.

From preselected conditions, the user can select "Bacteria", "Archaea", "Fungi", "Eukaryota", "no Fungi", "Fungi", "Viruses", "Bacteria", "Eukaryota", or "Archaea" to restrict the UniProt IDs in the subunit to those belonging to those taxonomy groups.

"Bacteria", "Archaea", "Fungi", "Bacteria", "Archaea", and "Fungi" selected categories may provide genome context (gene clusters/operons) useful for inferring functions.

The UniProt IDs also can be restricted to taxonomy categories within the Superkingdom, Kingdom, Phylum, Class, Order, Family, Genus, and Species ranks. Multiple conditions are combined to be a union of each other.

Preselected conditions: select a preset to auto populate

Add Taxonomy category

Length Filter

Job name: (required)

E-mail address:

You will be notified by e-mail when your submission has been processed.

Previous Jobs Families FASTA Accession IDs

Retrieve taxonomy for families.

The UniProt IDs for family members are identified in UniProtKB with a list of Pfam families, InterPro families, and/or Pfam clans.

Pfam and/or InterPro Families: 

45784 PF04055 PF06969 PF08497 PF12345 PF13186 PF16199 PF16881 PF19238 PF19288 PF19864

Family	Family Name	Full Size	UniRef90 Size	UniRef50 Size
IPR000385	MoaA_NiFB_PqqE_Fe-S-bd_CS	49,241	23,160	4,777
IPR001989	Radical_activat_CS	26,935	9,907	1,836
IPR002684	Biotin_synth/BioAB	27,640	9,880	1,004
IPR003698	Lipoyl_synth	39,047	13,924	1,318
IPR003739	Lys_aminomutase/Glu_NH3_mut	20,775	10,372	1,278
IPR004383	rRNA_Isu_MTrfase_RlmN/Cfr	39,944	15,429	1,455
IPR004558	Coprogen_oxidase_HemN	16,796	6,746	513
IPR004559	HemW-like	38,255	17,990	2,765
IPR005839	Methylthiotransferase	87,716	37,295	4,127
IPR005840	Ribosomal_S12_MeSTrfase_RimO	28,658	11,857	2,029
IPR005909	RaSEA	2,035	1,022	229
IPR005911	YhcC-like	11,293	4,505	502
IPR005980	Nase_CF_NiFB	2,647	1,104	82
IPR006463	MiaB_methiolase	35,615	13,649	723
IPR006466	MiaB-like_B	4,407	2,226	506
IPR006467	MiaB-like_C	17,077	8,216	1,089
IPR006638	Elp3/MiaA/NiFB-like_rSAM	446,282	212,389	36,535
IPR007197	rSAM	722,535	355,669	70,723
IPR010505	Mob_synth_C	38,361	16,635	1,682
IPR010722	BATS_dom	39,852	14,953	1,222
IPR010723	HemN_C	39,495	17,002	2,804
IPR011101	DUF5131	7,313	4,801	1,324
IPR011843	PQQ_synth_PqqE_bac	5,549	1,835	59

Fragment Option

UniProt designates a Sequence Status for each member. Complete if the encoding DNA sequence has both start and stop codons; Fragment if the start and/or stop codon is missing. Approximately 10% of the entries in UniProt are fragments.

Fragments: Check to exclude UniProt-defined fragments in the results. (default: off)

For the UniRef90 and UniRef50 databases, clusters are excluded if the cluster ID ("representative sequence") is a fragment.

UniProt IDs in UniRef90 and UniRef50 clusters with complete cluster IDs are removed from the clusters if they are fragments.

Filter by Taxonomy

This filter is applied to the UniProt IDs after they have been identified using the list of Pfam families, InterPro families, and/or Pfam clans. The remaining UniProt IDs are used to generate the sunburst.

From preselected conditions, the user can select "Bacteria, Archaea, Fungi", "Eukaryota, no Fungi", "Fungi", "Viruses", "Bacteria", "Eukaryota", or "Archaea" to restrict the UniProt IDs in the sunburst to these taxonomy groups.

"Bacteria, Archaea, Fungi", "Bacteria", "Archaea", and "Fungi" select organisms that may provide genome context (gene clusters/operons) useful for inferring functions.

The UniProt IDs also can be restricted to taxonomy categories within the Superkingdom, Kingdom, Phylum, Class, Order, Family, Genus, and Species ranks. Multiple conditions are combined to be a union of each other.

Preselected conditions:

Length Filter

Job name: (required) 

E-mail address: 

You will be notified by e-mail when your submission has been processed.



The results were available on the **DATASET COMPLETED** page. The **Taxonomy Sunburst** tab (below) is the interactive display that provides the numbers of UniProt, UniRef90 cluster, and UniRef50 cluster IDs (red arrow), downloads for IDs and FASTA sequences (blue arrow), and transfers of IDs to EFI-EST (green arrow) or EFI-GND viewer (magenta arrow).

DATASET COMPLETED

Submission Name: **IP91_RSS_All**

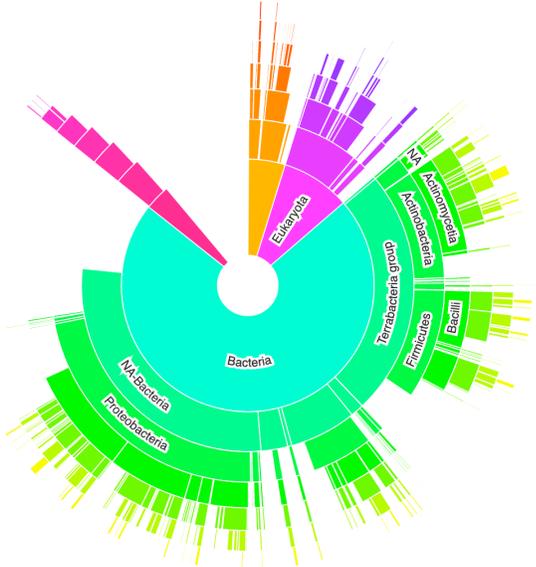
Dataset Summary
Taxonomy Sunburst
Length Histograms

The taxonomy distribution for the UniProt IDs identified as members of the input list of families is displayed.

The UniRef90 and UniRef50 clusters containing the UniProt IDs in the sunburst are identified using the lookup table provided by UniProt/UniRef. These UniRef90 and UniRef50 clusters may contain UniProt IDs from other families; in addition, the UniRef90 and UniRef50 clusters at a selected taxonomy category may contain UniProt IDs from other categories. This results from conflation of UniProt IDs in UniRef90 and UniRef50 clusters that share $\geq 90\%$ and $\geq 50\%$ sequence identity, respectively.

The numbers of UniProt IDs, UniRef90 cluster IDs, and UniRef50 cluster IDs for the selected category are displayed.

The sunburst is interactive, providing the ability to zoom to a selected taxonomy category by clicking on that category; clicking on the center circle will return the display to the next highest rank.



- Root
- Superkingdom
- Kingdom
- Phylum
- Class
- Order
- Family
- Genus
- Species

773,426 UNIPROT, 385,429 UNIREF90, 83,707 UNIREF50 IDS
←

Lists of UniProt, UniRef90, and UniRef50 IDs and FASTA-formatted sequences can be downloaded.

The UniProt, UniRef90, and UniRef50 IDs can be transferred to the Accession IDs option of EFI-EST to generate an SSN. The Accession IDs option provides both Filter by Family and Filter by Taxonomy that should be used to remove internal UniProt IDs from UniRef90 or UniRef50 clusters that are not members of the selected families and/or taxonomy category.

The lists also can be transferred to the GND-Viewer to obtain GND.

ID type: UniProt UniRef90 UniRef50

Prepare ID Download
Prepare FASTA Download
Transfer to EFI-EST
Transfer to EFI-GND Viewer

Complete Sequences. The **Taxonomy Sunburst** for the complete sequences was generated from UniProt Release 2022_04 by entering a list of 211 Pfam and InterPro families and/or domains (**Tutorial Table 1**) into the **Pfam and/or InterPro Families and/or Pfam Clans** box on the **Taxonomy Tool Families Option** page (red arrow), selecting **Fragment Option** to exclude fragments (blue arrow), entering the **Job name** (orange arrow) and an **E-mail address** (cyan arrow), and clicking **“Submit analysis”** (black arrow). The Tool provides the list of input families/domains, so the page is long (next page). The two boxed areas are enlarged on the following page to clearly show the input parameters.

Previous Jobs Families FASTA Accession IDs

Retrieve taxonomy for families.

The UniProt IDs for family members are identified in UniProtKB with a list of Pfam families, InterPro families, and/or Pfam clans.

Pfam and/or InterPro Families:

45784 PFD0405 PFD3693 PFD0847 PF12345 PF11386 PF16159 PF16881 PF19238 PF19288 PF19864

Family	Family Name	Full Size	UniProt50 Size	UniProt50/52
IPR000393	MoaA_NiB_Pfb_Pe-9-bd_CS	49,241	23,100	4.777
IPR019389	Radical_acti_ni_CS	26,935	9,907	1,936
IPR002684	Biotin_synthBiAB	27,640	9,880	1,004
IPR003688	Lipoyl_synth	35,047	13,824	1,318
IPR007739	Lys_monomerclass_Glu_NHS_int	20,775	10,372	1,278
IPR024283	rRNA_his_M7586e_RRNCD	39,844	15,429	1,455
IPR045858	Coproporphyrinogen_III_HemN	16,796	6,746	513
IPR045599	HemN-like	38,255	17,960	2,765
IPR005880	Methyltransferase	87,716	37,295	4,127
IPR005849	ribosomal_S12_M67Thiase_RfmO	28,658	11,957	2,029
IPR005959	AcSseA	2,035	1,022	329
IPR005911	YhcC-like	11,293	4,505	502
IPR005980	Nase_CF_NiB	2,647	1,104	82
IPR004643	MiB_mntrase	35,615	13,949	723
IPR004646	MiB-like_B	4,407	2,208	506
IPR004647	MiB-like_C	17,077	8,216	1,089
IPR006638	Ep33MaANiB-like_SAM	722,232	212,389	36,535
IPR027187	ISAM	448,220	355,669	79,723
IPR010504	Mab_synth_C	38,361	16,635	1,862
IPR010722	BATS_dom	39,652	14,953	1,222
IPR010723	HemC_C	39,496	17,002	2,804
IPR011101	DUF3131	7,313	4,801	1,524
IPR011483	PFD_0391_synth_PoqE_bac	6,246	1,835	58
IPR012837	NeuG	11,645	4,135	630
IPR012838	Pf1_1_activating	9,425	2,995	180
IPR012839	Cyprax_psd1_activase	16,374	1,841	1,127
IPR013483	SPASAM	33,719	14,394	976
IPR013704	LPF0313_N	12,479	4,760	404
IPR013843	Methyltransferase_N	91,463	39,689	5,468
IPR013817	rRNA_sph2000a-synth	8,013	4,099	502
IPR014191	Anaer_NiNtr_activator	584	303	27
IPR016431	Pyruvate_methylase_activator	16,215	11,197	1,144
IPR016771	Fe-S_CyRbase_SAM_TM048B_prl	62	39	6
IPR016779	ISAM_MSM02088	1,820	751	63
IPR016682	Emv1	58	32	3
IPR017200	PoqE-like	19,368	10,212	2,373
IPR017872	ML_4551-like	433	209	15
IPR017742	Deazaamino_synth	2,879	884	27
IPR017853	Hopanoyl_synth-asso_SAM_HSH	4,462	1,470	62
IPR017834	Hopanoyl_synth-asso_SAM_HpJ	1,463	448	14
IPR017839	CuR2_family	6,705	3,144	254
IPR019940	CoR1_family	6,240	2,738	154
IPR020052	Co_2_synth-act	20,361	8,403	490
IPR020812	Methyltransferase_CS	96,629	40,900	5,265
IPR022431	Cyclic_DiHPL_synthase_mprC	6,120	2,294	95
IPR022432	MprE	5,864	2,299	97
IPR022417	Lys_monomerclass	1,892	890	38
IPR022459	Lysine_ammomase	3,211	1,685	92
IPR022482	EpmB	5,760	2,170	180
IPR022881	rRNA_his_M7586e_Civ	258	54	2
IPR022882	SPASAM	13,008	5,108	563
IPR023454	ISAM_horseho	197,532	100,116	21,963
IPR023805	Unharaadentod_Epi-rel	1,028	422	9
IPR023807	PaqB6_mod_SAM	116	104	37
IPR023819	PaqB-mod_ISAM_AF0577	1,476	492	22
IPR023820	ISAM_GDI-asso	219	33	1
IPR023821	ISAM_Tad-asso	2,039	1,236	206
IPR023822	ISAM_Tad-asso_bac	438	178	22
IPR023826	ISAM_HemB	136	61	9
IPR023842	CHP0360_SAM	7,036	3,804	568
IPR023843	ISAM_PT0314	62	36	5
IPR023847	SupHase_maturase_SAM	26,568	13,357	5,204
IPR023903	7-CO-7-deazaGln_synth_pul_Cko	534	290	13
IPR023814	DNA_SAM_pul	8,822	3,282	228
IPR023890	Banyasucc_Synthase_activating	10	8	1
IPR023885	4Fe-4S-binding_SPASAM_dom	60,444	34,786	13,298
IPR023906	DH_ArchD_gsu_nukration	777	235	15
IPR023901	Pmp19_PyB	286	172	23
IPR023897	SpoP_PP_ase	1,810	628	20
IPR023904	Paq_SAM_mat_YyG	36	16	6
IPR023912	YjW_bact	2,077	431	55
IPR023913	hncI	1,476	492	22
IPR023930	NuJ1	269	128	5
IPR023969	CHP04072_B12-60/SAM	68	60	15
IPR023979	CHP04074_B12-60/SAM	212	108	15
IPR023980	CHP04073_B12-60/SAM	590	306	151
IPR023984	ISAM_mn1_1	1,314	927	397
IPR023992	HemC1_Synth_NuJ	1,137	491	8
IPR023993	TYW1_archaea	2,475	1,376	136

IPR023997	hncI	3,939	1,320	322
IPR024001	Cypr-act_mpr_SAM_mn1_CopM	251	127	67
IPR024007	FaFeHyd_mn1_HuG5	3,483	1,754	79
IPR024016	CHP04054_SAM	20	14	4
IPR024017	Paq_syl_SAM	45	6	2
IPR024018	CHP04052_SAM	200	159	14
IPR024021	FaFeHyd_HuE_SAM	3,733	2,075	217
IPR024023	ISAM_paired_HisB	1,143	625	125
IPR024025	SCIFF_SAM_maturase	2,738	1,232	79
IPR024032	ISAM_paired_HisC	1,002	526	142
IPR024177	Biotin_synthase	24,304	8,337	476
IPR024511	DUF3641	4,310	2,730	235
IPR024560	LPF0313_C	11,093	3,800	317
IPR024561	7-CO-7-deazaGln_synth-like	25,740	11,973	1,164
IPR024565	LAM_C_dom	6,845	3,760	332
IPR024632	Genoep_mn1_SAM	106	95	26
IPR024633	HuW	1,392	438	75
IPR024635	SAM_SPASM_FusB	1,094	895	82
IPR024644	ISAM_SAM_Some	66	62	7
IPR024646	SCM_SAM_SomeF	59	56	6
IPR024651	ISAM_ScysA	4,383	2,704	272
IPR024657	ISAMSPASM_pnt_GRRM_ysike	342	189	72
IPR024661	CDOX_mntr	139	93	26
IPR024644	SAM_m_lpo	429	175	16
IPR024607	SAM_OG-Bacter	267	131	67
IPR024612	ISAM_Coxv_yl	171	108	34
IPR024623	ISAM_cytoep	160	138	16
IPR024626	ISAM_Ethanolamin	18	18	5
IPR024629	MA_synthase	13	10	5
IPR024641	B12_SAM_TM0216	427	325	66
IPR024642	ISAM_syl1_2	44	1	1
IPR024642	rRNA_his_M7586e_RRNm	37,070	13,980	1,140
IPR027205	Lipoyl_synth-act	337	132	20
IPR027527	Lipoyl_synth_m1	285	79	14
IPR027593	B12_SAM-rlpo	442	144	8
IPR027564	HspH_B12_SAM	427	155	5
IPR027570	GenRSP_SAM	60	51	2
IPR027583	ISAM_ACDX	154	98	8
IPR027595	SAM_mntr_m1	167	146	41
IPR027596	Amnalkamkdam_rB	10,871	2,712	845
IPR027604	W_SAM_matur	461	350	101
IPR027608	SpoP_SASAC	300	177	83
IPR027609	ISAM_Cant1_Proteobac	2,527	501	21
IPR027611	ISAM_Cant1_omn	3,044	1,315	28
IPR027622	ISAM_Cyt10r	50	11	3
IPR027628	Pseudo_SAM_Halo	248	67	1
IPR027633	ISAM_NuJ2	269	186	5
IPR028001	Co_2_2_2_mn1_m1	178	96	2
IPR030881	B12_SAM_cot1r	65	41	2
IPR030884	Aha_Proteobactera	570	287	4
IPR030896	ISAM_AAD_1	580	364	19
IPR030905	CoR2_activ_SAM	564	199	16
IPR030915	ISAM_SiB1	33	10	3
IPR030933	NuN_nor_SAM	34	21	4
IPR030950	ISAM_PyoJ	100	106	59
IPR030968	B12_SAM_PyT	74	44	2
IPR030977	CoR1_CoR2CoR3	3,347	1,835	23
IPR030989	SAM_XyB	38	16	3
IPR031003	CoR2_PhpC_SAM	97	75	29
IPR031004	ISAM_TRad	1,610	626	16
IPR031015	ISAM_mn1_rch	702	29	4
IPR031012	ISAM_mn1_paired	902	101	4
IPR031014	ISAM_BiE	39	23	1
IPR031015	Paq_2_3_mn1_mntr	160	51	4
IPR031019	ISAM_mn1_rch	11	11	6
IPR031981	LAM_N	25,755	8,384	724
IPR032432	Radical_SAM_C	20,918	9,267	1,452
IPR033971	Amalmyr_nitrogenase	9	3	1
IPR033974	Diphymo_diphosphate_activase	3	3	3
IPR033975	TrpH-like	9	3	1
IPR033976	GntE-like	9	1	1
IPR034165	NiB_C	2,248	884	94
IPR034389	hncI-like	5	2	2
IPR034391	Cmo-like_SPASM_omnHsp	9,351	3,670	2,830
IPR034405	F4o	26,713	12,035	1,325
IPR034422	HydE/PyB-like	5,384	3,253	647
IPR034423	TrpHnCuHydE-like	10,800	4,469	454
IPR034436	hncI-like-like	11	9	1
IPR034438	4-Phe_dextransylase_activase	15	1	1
IPR034457	Organic_radical-activating	53,598	26,655	4,922
IPR034462	Banyasucc_synthase_activase	32	11	2
IPR034465	PrpA-like_activase	3,808	433	27
IPR034466	Methyltransferase_Class_B	21,110	22,736	7,891
IPR034471	GGDTMA_synthase	1,393	710	47
IPR034474	Methyltransferase_Class_D	5,445	3,148	797
IPR034479	hncI-like	842	477	11
IPR034480	hncI-like	344	262	37
IPR034485	Anaerobic_Cypr-act_maturase	862	365	75
IPR034491	Anaerobic_Su_sulfatase-maturase	5,937	1,729	148
IPR034492	Anaerobic_Cypr-act_maturase	66	11	4

IPR034498	Bacteriochlorophyll_C8_MT	47	53	1
IPR034505	Coproporphyrinogen-III_Lyase	68,641	31,743	5,880
IPR034508	Spectromycin_synthetase	8	7	5
IPR034514	TrpH-like	11	3	1
IPR034515	TrpH-like	8	4	2
IPR034519	TrpH-like	11	6	1
IPR034520	Fom3-like	5	2	1
IPR034530	HspH-like	3,569	1,910	158
IPR034531	Methylatn_of_lysineamycin	9	4	1
IPR034532	DnaB-like	57	53	11
IPR034534	Pyrimidine_methyltransferase	8	7	1
IPR034547	The118a_maturase	21	3	3
IPR034548	rRNA_sph2000a-synthase	6,265	3,358	768
IPR034557	TrpH_rRNA_METHtransferase	3,721	1,422	120
IPR034559	SpoP_PP_ase	537	327	20
IPR034560	SpoP_PP_ase	1,273	302	6
IPR034567	ELP3-like	7,115	3,099	398
IPR034585	Methyltransferase_N_f	91,072	39,494	5,362
IPR036861	ELP3	23,268	10,427	1,876
IPR040072	Methyltransferase_A	43,050	17,542	2,586
IPR040074	BactDPRAY19	9,101	3,899	362
IPR040081	CoxI-like	1	2	1
IPR040082	GenV-like	4	1	1
IPR040085	M0074-like	7,375	4,888	824
IPR040096	M0093-like	22,149	11,842	1,979
IPR040097	M0021-like	1,843	1,037	231
IPR040098	M0103-like	747	535	117
IPR041582	RimO_TRAM	25,468	10,202	1,260
IPR043573	Pu_mntr_SAM-like_N	4,919	2,855	513
IPR045587	CoR1CoR2CoR3	21,334	9,947	956
IPR045784	Radical_SAM_N2	11,865	6,446	1,460
PFD04055	Radical_SAM	672,681	327,815	62,860
PFD0699	HemC_C	39,495	17,002	2,804
PFD0697	Radical_SAM_N	12,479	4,760	404
PF12345	DUF3641	4,310	2,730	235
PF13186	SPASAM	47,292	26,993	9,956
PF16199	Radical_SAM_C	20,918	9,267	1,452
PF16881	LAM_N	25,755	8,384	724
PF18028	Radical_SAM_2	4,615	2,855	513
PF18028	CoR1_C	21,334	9,947	956
PF19864	Radical_SAM_N2	11,865	6,446	1,460
Total		4,055,068	1,696,246	336,149

The input format is a single family or comma-separated list of families. Families should be specified as PFDxxx (five digits), IPRxxxxxx (six digits) or CLxxxx (four digits) for Pfam clans.

Filter by Taxonomy can be used to remove UniProt IDs that do not match the specified taxonomy categories.

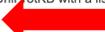
UniProt50 and UniProt52 cluster members that contain the UniProt IDs are retrieved from the UniProt50 and UniProt52 databases using the lookup table provided by UniProt/UniProt. Clusters for which the cluster ID (representative accession) matches the list of families are returned.

The numbers of UniProt IDs and both UniProt50 cluster

Previous Jobs Families FASTA Accession IDs

Retrieve taxonomy for families.

The UniProt IDs for family members are identified in UniProtKB with a list of Pfam families, InterPro families, and/or Pfam clans.

Pfam and/or InterPro Families: 

45784 PF04055 PF06969 PF08497 PF12345 PF13186 PF16199 PF16881 PF19238 PF19288 PF19864

Family	Family Name	Full Size	UniRef90 Size	UniRef50 Size
IPR000385	MoaA_NifB_PqqE_Fe-S-bd_CS	49,241	23,160	4,777
IPR001989	Radical_activat_CS	26,935	9,907	1,836
IPR002684	Biotin_synth/BioAB	27,640	9,880	1,004
IPR003698	Lipoyl_synth	39,047	13,924	1,318
IPR003739	Lys_aminomutase/Glu_NH3_mut	20,775	10,372	1,278
IPR004383	rRNA_Isu_MTrfase_RlmN/Cfr	39,944	15,429	1,455
IPR004558	Coprogen_oxidase_HemN	16,796	6,746	513
IPR004559	HemW-like	38,255	17,990	2,765
IPR005839	Methylthiotransferase	87,716	37,295	4,127
IPR005840	Ribosomal_S12_MeSTrfase_RimO	28,658	11,857	2,029
IPR005909	RaSEA	2,035	1,022	229
IPR005911	YhcC-like	11,293	4,505	502
IPR005980	Nase_CF_NifB	2,647	1,104	82
IPR006463	MiaB_methiolase	35,615	13,649	723
IPR006466	MiaB-like_B	4,407	2,226	506
IPR006467	MiaB-like_C	17,077	8,216	1,089
IPR006638	Elp3/MiaA/NifB-like_rSAM	446,282	212,389	36,535
IPR007197	rSAM	722,535	355,669	70,723
IPR010505	Mob_synth_C	38,361	16,635	1,682
IPR010722	BATS_dom	39,852	14,953	1,222
IPR010723	HemN_C	39,495	17,002	2,804
IPR011101	DUF5131	7,313	4,801	1,324
IPR011843	PQQ_synth_PqqE_bac	5,549	1,835	59

Fragment Option

UniProt designates a Sequence Status for each member. Complete if the encoding DNA sequence has both start and stop codons; Fragment if the start and/or stop codon is missing. Approximately 10% of the entries in UniProt are fragments.

Fragments: Check to exclude U ments in the results. (default: off)

For the UniRef90 and UniRef50 databases, clusters are excluded if the cluster ID ("representative sequence") is a fragment.

UniProt IDs in UniRef90 and UniRef50 clusters with complete cluster IDs are removed from the clusters if they are fragments.

Filter by Taxonomy

This filter is applied to the UniProt IDs after they have been identified using the list of Pfam families, InterPro families, and/or Pfam clans. The remaining UniProt IDs are used to generate the sunburst.

From preselected conditions, the user can select "Bacteria, Archaea, Fungi", "Eukaryota, no Fungi", "Fungi", "Viruses", "Bacteria", "Eukaryota", or "Archaea" to restrict the UniProt IDs in the sunburst to these taxonomy groups.

"Bacteria, Archaea, Fungi", "Bacteria", "Archaea", and "Fungi" select organisms that may provide genome context (gene clusters/operons) useful for inferring functions.

The UniProt IDs also can be restricted to taxonomy categories within the Superkingdom, Kingdom, Phylum, Class, Order, Family, Genus, and Species ranks. Multiple conditions are combined to be a union of each other.

Preselected conditions:

Add Taxonomy category

Length Filter

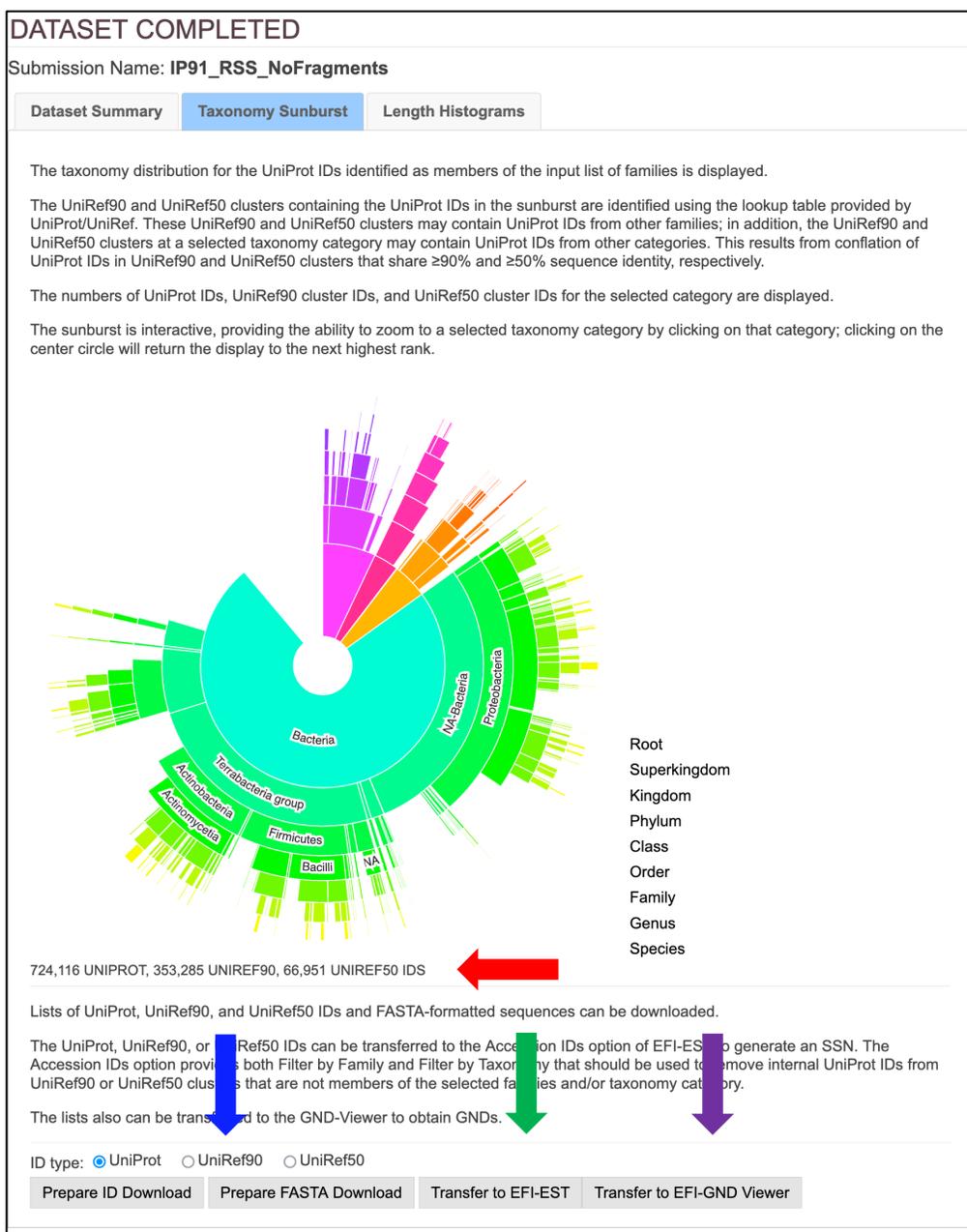
Job name: (required) 

E-mail address: 

You will be notified by e-mail when your submission has been processed.



The results were available on the **DATASET COMPLETED** page. The **Taxonomy Sunburst** tab (below) is the interactive display that provides the numbers of UniProt, UniRef90 cluster, and UniRef50 cluster IDs (red arrow), downloads for IDs and FASTA sequences (blue arrow), and transfers of IDs to EFI-EST (green arrow) or EFI-GND viewer (magenta arrow).



Complete Sequences, Minimum Length 140 Residues. The **Taxonomy Sunburst** for the complete sequences with a minimum length of 140 residues (“full-length” sequences) was generated by entering a list of 211 Pfam and InterPro families and/or domains (**Tutorial Table 1**) into the **Pfam and/or InterPro Families and/or Pfam Clans** box on the **Taxonomy Tool Families Option** page (red arrow), selecting **Fragment Option** to exclude fragments (blue arrow), entering 650 for the **Minimum Length** in the **Length Filter** (green arrow), entering the **Job name** (orange arrow) and an **E-mail address** (cyan arrow), and clicking “**Submit analysis**” (black arrow). The Tool provides the list of input families/domains, so the page is long (next page). The two boxed areas are enlarged on the following page to clearly show the input parameters.

Previous Jobs	Families	FASTA	Accession IDs
Retrieve taxonomy for families.			
The UniProt IDs for family members are identified in UniProt3 with a list of Plant families, InterPro families, and/or Plant clans.			
Filter and/or Filter Families and/or Plant clans			
43764 PF04055 PF06969 PF06487 PF12345 PF11308 PF16199 PF16881 PF19238 PF19288 PF19864			
Family	Family Name	Full Size	UniProt3 Size
IPR000861	MaA_MIP_MIPc_Fe-Sul_CS	49,241	23,160
IPR019689	Radical_Activ_C5	26,535	9,507
IPR020684	Biotin_activ_BiA8	27,840	9,880
IPR030998	LipoL_Synth	39,047	13,924
IPR007338	Lys_aminomutase_C1c_NH2_mut	20,775	10,372
IPR043483	RNA_hu_Mtase_RhmCN	39,944	16,429
IPR040508	Coprogen_citase_HemN	16,786	6,746
IPR040509	HemK-like	38,205	17,990
IPR000862	Methyltransferase	87,716	37,295
IPR006640	Ribonuclease_S12_MotIFase_RhmC	26,658	11,857
IPR009098	RaSEA	2,035	1,022
IPR009911	YhcC-like	11,203	4,505
IPR009860	Nase_C1_MIB	2,847	1,534
IPR006643	MaB_methase	36,615	13,649
IPR006646	MaB-like_B	4,077	2,205
IPR006476	MaB-like_C	17,407	8,216
IPR006652	ES33MaA/NBlike_GSAM	446,282	212,389
IPR007197	rSAM	722,535	305,689
IPR110005	Ma3_synth_C	38,381	16,635
IPR107322	BATS_dom	39,862	14,933
IPR107323	Ma3_synth_N	39,862	14,933
IPR117171	SOV371	7,172	2,707
IPR118143	PO2_synth_PoE2_bac	5,549	1,835
IPR127226	ThiH	6,704	2,297
IPR128257	hNH3	11,645	4,136
IPR112838	PFL1_activating	9,425	2,995
IPR112839	Organic_nitro_activase	16,374	5,941
IPR113403	MaA	33,719	14,394
IPR13702	UPF0313_N	12,879	4,760
IPR118448	Methyltransferase_N	91,463	39,689
IPR139117	RNA_nitrosylase_synth	6,013	4,099
IPR141919	Anaer_RNR_activator	594	303
IPR159423	Pyruvate_kinase_activator_gnt	16,215	11,167
IPR167771	Fes_Oxidase_rSAM_TM0486_gnt	62	30
IPR167779	rSAM_MSME00568	1,825	751
IPR168902	DxliI	98	32
IPR170202	PaE-like	10,368	10,212
IPR119732	MA_451-like	433	209
IPR117442	Deoxyguanine_synth	2,879	884
IPR117833	Hopanoid_synth_activase_rSAM_HprH	4,542	1,470
IPR117834	Hopanoid_synth_activase_rSAM_HprU	1,453	448
IPR119938	CuI2_family	6,705	3,144
IPR119940	CuI2_family	6,240	2,738
IPR200050	FO_synthase_sul2	20,381	8,463
IPR200102	Methyltransferase_CS	96,239	40,899
IPR202431	Cyclic_DHPL_synthase_hmpC	6,100	2,294
IPR202432	MayE	5,884	2,259
IPR202447	Lys_aminomutase_nrl	1,692	880
IPR202450	Lysine_aminomutase	3,211	1,685
IPR202462	ExpB	6,760	2,170
IPR228818	RNA_hu_Mtasea_C9	256	54
IPR229246	UPF0313	13,028	5,108
IPR229432	rSAM_Nasease	197,532	102,156
IPR229605	Urncharakterised_Spl-nrl	1,028	422
IPR229607	Papides_mod_rSAM	116	104
IPR229819	Pap-mod_rSAM_AF0577	682	468
IPR229820	rSAM_GCL_activase	219	33
IPR229821	rSAM_Nad-asso	2,039	1,236
IPR229822	rSAM_Tad-asso	438	178
IPR229828	rSAM_HemB	135	61
IPR229830	CHP0360_rSAM	7,935	3,804
IPR229833	rSAM_FT01314	82	36
IPR229867	Sulphatase_maturase_rSAM	26,888	13,657
IPR229868	7-CD-7-oxaaz-Gua_synth_hu_C1c	534	250
IPR229871	DM1_rSAM_sul	8,922	3,282
IPR229880	Benzoylase_Synthase_activating	10	8
IPR229885	4Fe4S-binding_SPMAS_dom	60,484	34,786
IPR229886	QH-AnhD1_glu_maturase	777	235
IPR229891	Pyruvate_PyB	286	172
IPR229897	Spore_PP_yasease	1,810	628
IPR229904	Pap_rSAM_mal_YuYD	36	16
IPR229912	Yjw1_bact	2,077	431
IPR229913	MAC	1,478	482
IPR229920	Nu1	269	128
IPR229969	CHP04072_B12-bidSAM	88	40
IPR229979	CHP0414_B12-bidSAM	212	108
IPR229980	CHP0412_B12-bidSAM	910	558
IPR229984	rSAM_ash_1	1,314	907
IPR229992	HemO1_synth_Nu1	1,137	491
IPR229993	TYW1_activase	2,475	1,376

IPR023995	HemZ	3,639	1,607
IPR024001	Cys-H1_Lys_rSAM_mal_CupM	231	130
IPR024007	Fe-Fe-H1_mal_HuG	3,483	1,754
IPR024016	CHP04054_rSAM	20	14
IPR024017	Pap_syl_rSAM	45	6
IPR024021	CHP04052_rSAM	320	159
IPR024021	Fe-Fe-H1_HuE_rSAM	3,733	2,075
IPR024023	rSAM_gated_HuB	1,143	625
IPR024025	SOFF_rSAM_maturase	2,738	1,202
IPR024032	rSAM_gated_HuC	1,002	526
IPR024177	Biotin_synthase	24,304	8,337
IPR024251	DUF3641	4,310	2,730
IPR024260	SOFF_rSAM_maturase	11,093	3,800
IPR024261	T-CD-7-oxaaz-guanine_synth-like	25,740	11,573
IPR024265	LAM_C_dom	6,845	3,760
IPR024322	Genop_mal_rSAM	106	95
IPR024332	HuWf	1,392	438
IPR024344	rSAM_SPMAS_FuB	1,394	895
IPR024344	rSAM_SPMAS_SomE	66	62
IPR024346	rSAM_SPMAS_SomF	99	96
IPR024351	rSAM_SPMAS	4,363	2,794
IPR024357	rSAM_SPMAS_gnt_ORRM_system	342	189
IPR024360	COXX_natur	139	93
IPR024604	rSAM_w_lipo	429	175
IPR024607	SAM_SPG-Beater	207	131
IPR024612	rSAM_Spm_Z1	171	108
IPR024623	rSAM_nobpam	190	138
IPR024626	rSAM_Fluorin	18	18
IPR024629	MA_synthase	13	10
IPR024641	B12_SAM_Nu274	437	267
IPR024642	rSAM_nu1_3	129	44
IPR024748	RNA_Mtasea_RhN	37,070	13,880
IPR027928	LipoL_synth_nrlpt	337	132
IPR027927	LipoL_synth_nrl	265	79
IPR027958	B12_rSAM_nlgo	443	144
IPR027964	HprH_B12_rSAM	427	155
IPR027970	MaRSP_rSAM	160	51
IPR027971	rSAM_nu20	167	58
IPR027986	rSAM_metal_mer	107	146
IPR027986	ArmoMeroGsm_rS	10,871	7,712
IPR027990	W_SAM_natur	481	300
IPR027992	Spore_SPMAS	177	107
IPR027993	rSAM_QueE_Proteobac	2,627	501
IPR027993	rSAM_QueE_gama	3,544	1,315
IPR027993	rSAM_QueE_CuI2	100	11
IPR027993	Fraxido_SAM_HuB	248	87
IPR027993	rSAM_nu12	399	186
IPR030001	Ox_2_3_NH2_mut	266	178
IPR030037	B12_rSAM_nrlc1	95	41
IPR030096	ANA_Pyruvate_activase	570	287
IPR030096	rSAM_ANO2_hemab	590	364
IPR030096	CuI2_activ_rSAM	564	190
IPR030096	rSAM_SMB	33	10
IPR030096	hml_yuo_rSAM	34	21
IPR030096	rSAM_PoY0	150	59
IPR030096	B12_rSAM_lpo_MT	74	46
IPR030097	QueE_CuI2CivC	3,947	1,835
IPR030098	rSAM_Synth	38	16
IPR031003	BuG_PhrG_rSAM	87	75
IPR031004	rSAM_Y8AB	1,610	526
IPR031010	rSAM_mol_garB	702	29
IPR031012	rSAM_mol_garB	902	101
IPR031014	rSAM_BHE	39	23
IPR031015	Arg_2_3_nH2_mut	160	51
IPR031019	rSAM_rSAM_C_nrl	11	11
IPR031089	LAM_N	25,755	8,384
IPR032432	Radical_SAM_C	20,918	9,267
IPR033971	Avanymon_spmomase	9	3
IPR033974	Glyoxylate_dehydratase_activase	33	5
IPR033975	ThpH-like	9	3
IPR033976	GHE-like	9	1
IPR034165	NIB_C	2,248	884
IPR034386	BvH-like	3	2
IPR034391	CuI2-like_SPMAS_containing	9,381	6,770
IPR034405	F4D2	26,713	12,035
IPR034422	HyeE/PvB-like	5,384	3,253
IPR034428	ThiHnCLHyG-like	10,820	4,489
IPR034438	NuChN-like	11	9
IPR034438	4-Fe4S-binding_activase	15	1
IPR034457	Organic_nitro_activating	53,586	26,655
IPR034462	Benzoylase_synthase_activase	21	4
IPR034465	Pyruvate_Synthase_activase	3,836	433
IPR034466	Methyltransferase_Class_B	32,110	22,736
IPR034471	GOOTAM_synthase	1,393	710
IPR034474	Methyltransferase_Class_D	6,445	3,184
IPR034475	ANIC-like	942	477
IPR034480	Heme_carbon_yase-like	344	262
IPR034485	Anaerobic_Cys-type_sulfatase-m	662	365
IPR034491	Anaerobic_Ser_sulfatase-maturase	5,937	1,729

IPR034485	Anaerobic_Cys-type_sulfatase-m	662	365
IPR034491	Anaerobic_Ser_sulfatase-maturase	5,937	1,729
IPR034497	Bacteriochlorophyll_C12_MT	56	37
IPR034498	Bacteriochlorophyll_C8_M1	47	33
IPR034505	Coproporphyrinogen-III_activase	68,641	31,743
IPR034508	Speitricin_synthase	8	7
IPR034514	ThxK-like	11	3
IPR034515	ThxL-like	8	4
IPR034519	Turb-like	11	6
IPR034529	FonH-like	5	2
IPR034530	HprP-like	3,969	1,910
IPR034531	Methylation_of_lysine	9	4
IPR034532	Oxid-like	97	53
IPR034534	Pyridoxal_methyltransferase	8	7
IPR034547	Thi1866_maturase	21	3
IPR034556	RNA_ributyltransferase	6,285	3,336
IPR034557	Thra_rRNA_MeItransferase	3,721	1,422
IPR034559	Spore_PP_yasease	637	327
IPR034560	Spore_PP_yasease_bact	1,273	302
IPR034567	ELP3-like	7,115	3,099
IPR034570	Methyltransferase_N_f	91,272	39,484
IPR034581	ELP3	25,268	10,427
IPR034582	Methyltransferase_A	43,050	17,542
IPR040074	BidPSA/Yjw	9,191	3,899
IPR040081	Cnd-like	1	2
IPR040082	Gnk-like	4	1
IPR040085	MJ027-like	7,375	4,868
IPR040095	MJ063-like	22,140	11,542
IPR040097	MJ021-like	1,543	1,037
IPR040098	MJ0103-like	477	535
IPR041582	RhD_THAM	25,468	10,202
IPR045375	Pol_radical_SAM-like_N	4,615	2,655
IPR045687	CofM/MuG-like_C	21,334	9,047
IPR045784	Radical_SAM_N2	11,865	6,448
PF04055	Radical_SAM	672,881	327,815
PF06969	HemN_C	39,495	17,002
PF06987	Radical_SAM_N	12,479	4,760
PF12345	DUF3641	4,310	2,730
PF13186	SPASM	47,292	26,903
PF16199	Radical_SAM_C	20,918	9,267
PF16881	LAM_N	25,755	8,384
PF19238	Radical_SAM_2	4,615	2,655
PF19288	CofL_C	21,334	9,047
PF19864	Radical_SAM_N2	11,865	6,448
Total		4,055,668	1,899,246
Total Completed		4,055,668	336,149

The input format is a single family or complex/space separated list of families. Families should be specified as PFxxxx (five digits), PFXxxxx (six digits) or CXXXX (four digits) for Plant clans.

Filter by Taxonomy can be used to remove UniProt IDs that do not match the specified taxonomy categories.

The remaining UniProt IDs are used to generate the surburst.

UniProt3 and UniProt4 clusters that contain the UniProt IDs are retrieved from the UniProt3 and UniProt4 databases using the UniProt IDs provided by UniProt3/UniProt4. Clusters for which the cluster ID (representative sequence) matches the list of families are retained.

The numbers of UniProt IDs and both UniProt3 cluster and UniProt4 cluster IDs are displayed on the surburst; the UniProt IDs and both UniProt3 cluster and UniProt4 cluster IDs are available for download and/or transfer to the Accession ID option (Option D) of EPI-EST to generate SSAs.

If the lists of UniProt3 or UniProt4 cluster IDs are used to generate SSAs with the Accession IDs option (Option D) of EPI-EST, the lists should (eventually) be filtered with the same list of families (Filter by Family) and any specified taxonomy categories (Filter by Taxonomy) used to generate the lists.

This filtering removes the UniProt3 and UniProt4 clusters with cluster IDs ("representative sequence") or internal UniProt IDs that do not match the specified taxonomy categories.

Fragment Option

UniProt designates a Sequence Status for each member. Complete if the encoding DNA sequence has both start and stop codons. Fragment if the start and/or stop codon is missing. Approximately 10% of the entries in UniProt are fragments.

Fragment: Check to exclude UniProt-defined fragments in the results. (default: off)

For the UniProt3 and UniProt4 databases, clusters are excluded if the cluster ID ("representative sequence") is a fragment.

UniProt IDs in UniProt3 and UniProt4 clusters with complete cluster IDs are removed from the clusters if they are fragments.

Filter by Taxonomy

Length Filter

Minimum Length:

Previous Jobs Families FASTA Accession IDs

Retrieve taxonomy for families.

The UniProt IDs for family members are identified in UniProtKB with a list of Pfam families, InterPro families, and/or Pfam clans.

Pfam and/or InterPro Families and/or Pfam clans: 

45784 PF04055 PF06969 PF08497 PF12345 PF13186 PF16199 PF16881 PF19238 PF19288 PF19864

Family	Family Name	Full Size	UniRef90 Size	UniRef50 Size
IPR000385	MoaA_NifB_PqqE_Fe-S-bd_CS	49,241	23,160	4,777
IPR001989	Radical_activat_CS	26,935	9,907	1,836
IPR002684	Biotin_synth/BioAB	27,640	9,880	1,004
IPR003698	Lipoyl_synth	39,047	13,924	1,318
IPR003739	Lys_aminomutase/Glu_NH3_mut	20,775	10,372	1,278
IPR004383	rRNA_Isu_MTrfase_RimN/Cfr	39,944	15,429	1,455
IPR004558	Coprogen_oxidase_HemN	16,796	6,746	513
IPR004559	HemW-like	38,255	17,990	2,765
IPR005839	Methylthiotransferase	87,716	37,295	4,127
IPR005840	Ribosomal_S12_MeSTrfase_RimO	28,658	11,857	2,029
IPR005909	RaSEA	2,035	1,022	229
IPR005911	YhcC-like	11,293	4,505	502
IPR005980	Nase_CF_NifB	2,647	1,104	82
IPR006463	MiaB_methiolase	35,615	13,649	723
IPR006466	MiaB-like_B	4,407	2,226	506
IPR006467	MiaB-like_C	17,077	8,216	1,089
IPR006638	Elp3/MiaA/NifB-like_rSAM	446,282	212,389	36,535
IPR007197	rSAM	722,535	355,669	70,723
IPR010505	Mob_synth_C	38,361	16,635	1,682
IPR010722	BATS_dom	39,852	14,953	1,222

Fragment Option

UniProt designates a Sequence Status for each member: Complete if the encoding DNA sequence has both start and stop codons; Fragment if the start and/or stop codon is missing. Approximately 10% of the entries in UniProt are fragments.

Fragments: Check to  refined fragments in the results. (default: off)

For the UniRef90 and UniRef50 databases, clusters are excluded if the cluster ID ("representative sequence") is a fragment.

UniProt IDs in UniRef90 and UniRef50 clusters with complete cluster IDs are removed from the clusters if they are fragments.

Filter by Taxonomy

Length Filter

Minimum Length: 

Maximum Length:

Job name: (required) 

E-mail address: 

You will be notified by e-mail when your submission has been processed.



UniRef50 Cluster SSN

The UniRef50 cluster SSN was generated for the complete UniRef50 cluster entries in the RSS using the **EFI-EST Families Option** by inserting the list of 211 Pfam and InterPro families and/or domains (**Tutorial Table 1**) into the **Pfam and/or InterPro Families and/or Pfam Clans** box (red arrow) and **UniRef50 cluster IDs** (blue arrow), checking the **Fragments** box in the **Fragment Option** to exclude fragments (green), entering the **Job name** (orange arrow) and an **E-mail address** (cyan arrow), and clicking **“Submit analysis”** (black arrow).

Previous Jobs Sequence BLAST **Families** FASTA Accession IDs SSN Utilities

Generate a SSN for a protein family.

The members of the input Pfam families, InterPro families, and/or Pfam clans are selected from the UniProt, UniRef90, or UniRef50 database.

Pfam and/or InterPro Families and/or Pfam clans: ←

45784 PF04055 PF06969 PF08497 PF12345 PF13186 PF16199 PF16881 PF19238 PF19288 PF19864

Use UniRef50 cluster ID sequences (UniProt is default). ←

Family	Family Name	Full Size	UniRef90 Size	UniRef50 Size
IPR000385	MoaA_NiFB_PqqE_Fe-S-bd_CS	49,241	23,160	4,777
IPR001989	Radical_activat_CS	26,935	9,907	1,836
IPR002684	Biotin_synth/BioAB	27,640	9,880	1,004
IPR003698	Lipoyl_synth	39,047	13,924	1,318
IPR003739	Lys_aminomutase/Glu_NH3_mut	20,775	10,372	1,278
IPR004383	rRNA_Isu_MTrfase_RlmN/Cfr	39,944	15,429	1,455
IPR004558	Coprogen_oxidase_HemN	16,796	6,746	513
IPR004559	HemW-like	38,255	17,990	2,765
IPR005839	Methylthioesterase	87,716	37,295	4,127
IPR005840	Ribosomal_S12_MeSTrfase_RimO	28,658	11,857	2,029
IPR005909	RaSEA	2,035	1,022	229
IPR005911	YhcC-like	11,293	4,505	502
IPR005980	Nase_CF_NiFB	2,647	1,104	82
IPR006463	MiaB_methylase	35,615	13,649	723
IPR006466	MiaB-like_B	4,407	2,226	506
IPR006467	MiaB-like_C	17,077	8,216	1,089
IPR006638	Elp3/MiaANiFB-like_rSAM	446,282	212,389	36,535
IPR007197	rSAM	722,535	355,669	70,723
IPR010505	Mob_synth_C	38,361	16,635	1,682
IPR010722	BATS_dom	39,852	14,953	1,222
IPR010723	HemN_C	39,495	17,002	2,804
IPR011101	DUF5131	7,313	4,801	1,324
IPR011843	PQQ_synth_PqqE_bac	5,549	1,835	59
IPR012726	ThiH	6,704	2,297	125
IPR012837	NrdG	11,645	4,135	630
IPR012838	PFL1_activating	9,425	2,995	160
IPR012839	Organic_radical_activase	16,374	5,941	1,127
IPR013483	MoaA	33,719	14,394	976

Fragment Option

UniProt designates a Sequence Status for each member. Complete if the encoding DNA sequence has both start and stop codons; Fragment if the start and/or stop codon is missing. Approximately 10% of the entries in UniProt are fragments.

Fragments: Check to exclude UniProt results. (default: off) ←

For the UniRef90 and UniRef50 databases, clusters are excluded if the cluster ID ("representative sequence") is a fragment.

UniProt IDs in UniRef90 and UniRef50 clusters with complete cluster IDs are removed from the clusters if they are fragments.

Filter by Taxonomy

From preselected conditions, the user can select "Bacteria, Archaea, Fungi", "Eukaryota, no Fungi", "Fungi", "Viruses", "Bacteria", "Eukaryota", or "Archaea" to restrict the retrieved sequences to these taxonomy groups.

"Bacteria, Archaea, Fungi", "Bacteria", "Archaea", and "Fungi" select organisms that may provide genome context (gene clusters/operons) useful for inferring functions.

The retrieved sequences also can be restricted to taxonomy categories within the SuperKingdom, Kingdom, Phylum, Class, Order, Family, Genus, and Species ranks. Multiple conditions are combined to be a union of each other.

The sequences from the UniRef90 and UniRef50 databases are the UniRef90 and UniRef50 clusters for which the cluster ID ("representative sequence") matches the specified taxonomy categories. The UniProt members in these clusters that do not match the specified taxonomy categories are removed from the cluster.

Preselected conditions: -- select a preset to auto populate --

Add Taxonomy category

Protein Family Size Options

Family Domain Boundary Option

SSN Edge Calculation Option

Job name: (required) ←

E-mail address: ←

You will be notified by e-mail when your submission has been processed.

←

The SSN was finalized on the **SSN Finalization** tab of the **DATASET COMPLETED** page using 11 as the **Alignment Score Threshold** (orange arrow) and 140 residues as the **Minimum** in the **Sequence Length Restriction** to remove truncated sequences (cyan arrow), entering the **Network (SSN) name** (brown arrow), and clicking “**Create SSN**” (black arrow). We previously determined that the members of the anaerobic ribonucleotide reductase activating enzyme family have the shortest sequences (≥ 140 residues) [2]).

DATASET COMPLETED

Submission Name: **IP91_RSS_UniRef50_NoFragments**

A minimum sequence similarity threshold that specifies the sequence pairs connected by edges is needed to generate the SSN. This threshold also determines the segregation of proteins into clusters. The threshold is applied to the edges in the SSN using the alignment score, an edge node attribute that is a measure of the similarity between sequence pairs.

Dataset Summary
Taxonomy Sunburst
Dataset Analysis
SSN Finalization

This tab is used to specify the minimum "Alignment Score Threshold" (that is a measure of the minimum sequence similarity threshold) for drawing the edges that connect the proteins (nodes) in the SSN.

Alignment Score Threshold: ?

This value corresponds to the lower limit for which an edge will be present in the SSN. The alignment score is similar in magnitude to the negative base-10 logarithm of a BLAST e-value.

▾ **Sequence Length Restriction Options**

Allows restriction of sequences in the generated SSN based on their length. ?

Minimum: (default: 0)

Maximum: (default: 50000)

▸ **Filter by Taxonomy**

▸ **Neighborhood Connectivity**

▸ **Fragment Option**

Network name: This name will be used in Cytoscape.

You will be notified by e-mail when the SSN is ready for download.

Create SSN

The **Network Files** tab of the **DOWNLOAD NETWORK FILES** page provided the xgmml file for the **Full (SSN) Network** (red arrow; all UniRef50 cluster nodes and edges) as well as the xgmml files for **Representative Node Networks** that conflate the UniProt nodes based on percent identity (blue arrow). The xgmml files for the full SSN were downloaded, opened with Cytoscape 3.9.1, and displayed with the Prefuse Force Directed layout; the nodes were colored according to the Structure-Function Linkage Database (SFLD) subgroups [2, 3].

DOWNLOAD NETWORK FILES

Submission Name: IP91_RSS_UniRef50_NoFragments
Network Name: IP91_RSS_UniRef50_NoFragments_Minlen140_AS11

SSN Overview Network Files

Please cite your use of the EFI tools:

Rémi Zallot, Nils Oberg, and John A. Gerit, The EFI Web Resource for Genomic Enzymology Tools: Leveraging Protein, Genome, and Metagenome Databases to Discover Novel Enzymes and Metabolic Pathways. Biochemistry 2019 58 (41), 4169-4182. <https://doi.org/10.1021/acs.biochem.9b00735>

The panels below provide files for full and representative node SSNs for download with the indicated numbers of nodes and edges. As an approximate guide, SSNs with ~2M edges can be opened with 16 GB RAM, ~5M edges can be opened with 32 GB RAM, ~10M edges can be opened with 64 GB RAM, ~20M edges can be opened with 128 GB RAM, ~40M edges can be opened with 256 GB RAM, and ~120M edges can be opened with 768 GB RAM.

Files may be transferred to the Genome Neighborhood Tool (GNT), the Color SSN utility, the Cluster Analysis utility, or the Neighborhood Connectivity utility.

Full Network ← ?

Each node in the network represents a single protein sequence.

	# Nodes	# Edges	
Download ZIP	63,359	65,098,917	Transfer To: ▼

Representative Node Networks ← ?

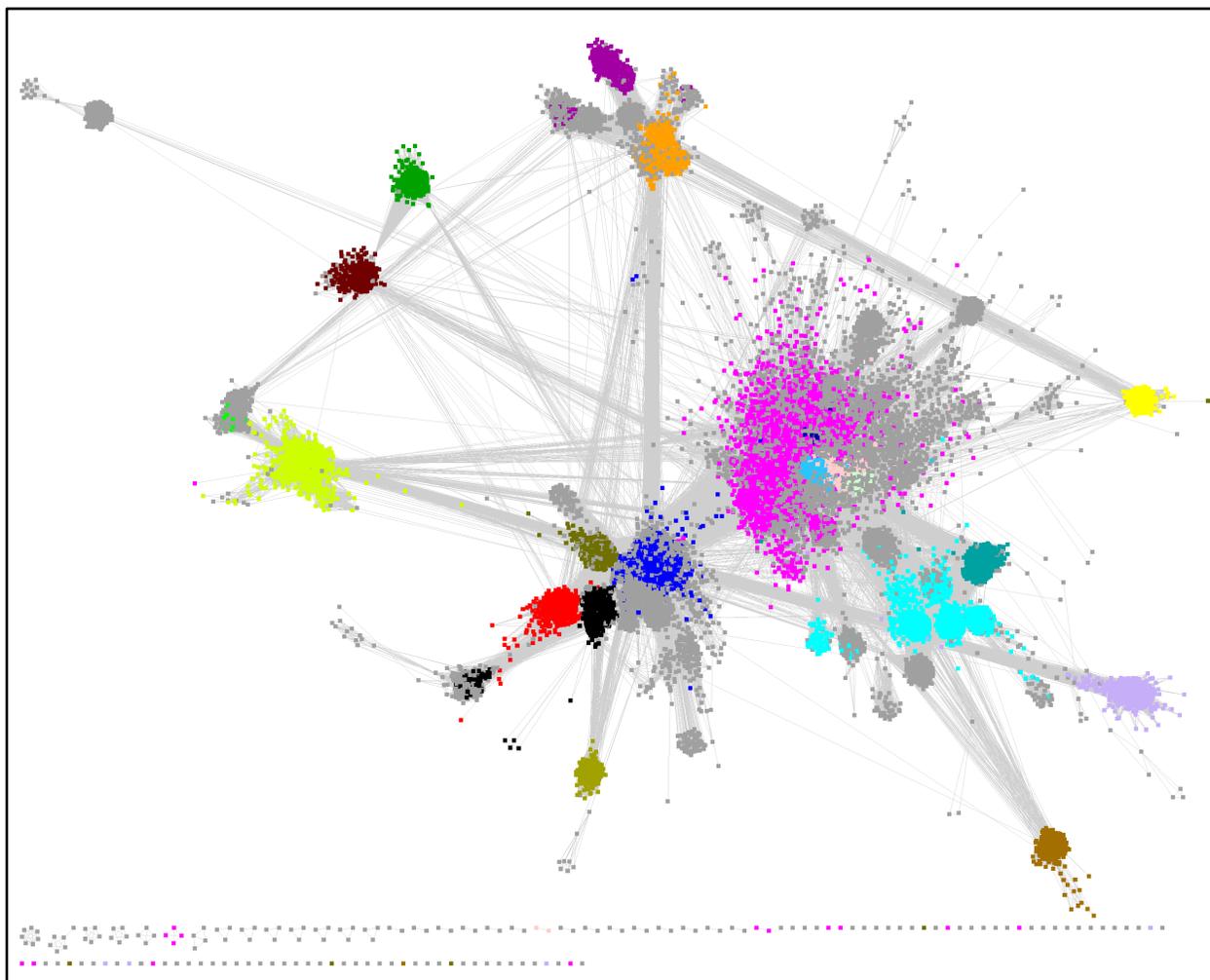
In representative node (RepNode) networks, each node in the network represents a collection of proteins grouped according to percent identity. For example, for a 75% identity RepNode network, all connected sequences that share 75% or more identity are grouped into a single node (meta node). Sequences are collapsed together to reduce the overall number of nodes, making for less complicated networks easier to load in Cytoscape.

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

	% ID	# Nodes	# Edges	
Download ZIP	100	63,359	65,098,917	Transfer To: ▼
Download ZIP	95	63,352	65,091,388	Transfer To: ▼
Download ZIP	90	63,346	65,080,111	Transfer To: ▼
Download ZIP	85	63,280	64,963,179	Transfer To: ▼
Download ZIP	80	63,200	64,838,286	Transfer To: ▼
Download ZIP	75	63,137	64,742,434	Transfer To: ▼
Download ZIP	70	63,069	64,623,644	Transfer To: ▼
Download ZIP	65	63,002	64,478,524	Transfer To: ▼
Download ZIP	60	62,898	64,282,426	Transfer To: ▼
Download ZIP	55	62,746	64,021,976	Transfer To: ▼
Download ZIP	50	62,281	63,194,549	Transfer To: ▼
Download ZIP	45	60,324	59,506,667	Transfer To: ▼
Download ZIP	40	55,995	50,990,536	Transfer To: ▼

Download Network Statistics as Table

New to Cytoscape?



UniRef50 Cluster SSN for the RS Superfamily. The UniRef50 cluster SSN for the RS superfamily was generated using a minimum length of 140 residues and an alignment score threshold of 11, opened with Cytoscape 3.9.1, and displayed with the Prefuse Force Directed layout using a Mac Pro computer with 1.5TB RAM. The nodes are colored according to the subgroups defined by the Structure-Function Linkage Database (SFLD) [2, 3]. The SSN contains 63,359 nodes and 65,098,917 edges.

UniRef90 Cluster SSN

The UniRef90 cluster SSN was generated for the complete UniRef90 cluster entries in the RSS using the **EFI-EST Families Option** by specifying the list of 211 Pfam and InterPro families and/or domains (**Tutorial Table 1**) into the **Pfam and/or InterPro Families and/or Pfam Clans** box (red arrow) and **UniRef90 cluster IDs** (blue arrow), checking the **Fragments** box in the **Fragment Option** to exclude fragments (green arrow), entering the **Job name** (orange arrow) and an **E-mail address** (cyan arrow), and clicking **“Submit analysis”** (black arrow).

Previous Jobs Sequence BLAST **Families** FASTA Accession IDs SSN Utilities

Generate a SSN for a protein family.

The members of the input Pfam families, InterPro families, and/or Pfam clans are selected from the UniProt, UniRef90, or UniRef50 database.

Pfam and/or InterPro Families and/or Pfam clans: ←

45784 PF04055 PF06969 PF08497 PF12345 PF13186 PF16199 PF16881 PF19238 PF19288 PF19864

Use UniRef90 cluster ID sequences (UniProt is default). ←

Family	Family Name	Full Size	UniRef90 Size	UniRef50 Size
IPR000385	MoaA_NiFB_PqqE_Fe-S-bd_CS	49,241	23,160	4,777
IPR001989	Radical_activat_CS	26,935	9,907	1,836
IPR002684	Biotin_synth/BioAB	27,640	9,880	1,004
IPR003698	LipoY_synth	39,047	13,924	1,318
IPR003739	Lys_aminomutase/Glu_NH3_mut	20,775	10,372	1,278
IPR004383	rRNA_Isu_MTrfase_RlmN/Cfr	39,944	15,429	1,455
IPR004558	Coprogen_oxidase_HemN	16,796	6,746	513
IPR004559	HemW-like	38,255	17,990	2,765
IPR005839	Methylthioesterase	87,716	37,295	4,127
IPR005840	Ribosomal_S12_MeSTrfase_RimO	28,658	11,857	2,029
IPR005909	RaSEA	2,035	1,022	229
IPR005911	YhcC-like	11,293	4,505	502
IPR005980	Nase_CF_NiFB	2,647	1,104	82
IPR006463	MiaB_methylase	35,615	13,649	723
IPR006466	MiaB-like_B	4,407	2,226	506
IPR006467	MiaB-like_C	17,077	8,216	1,089
IPR006638	Elp3/MiaANiFB-like_rSAM	446,282	212,389	36,535
IPR007197	rSAM	722,535	355,669	70,723
IPR010505	Mob_synth_C	38,361	16,635	1,682
IPR010722	BATS_dom	39,852	14,953	1,222
IPR010723	HemN_C	39,495	17,002	2,804
IPR011101	DUF5131	7,313	4,801	1,324
IPR011843	PQQ_synth_PqqE_bac	5,549	1,835	59
IPR012726	ThiH	6,704	2,297	125
IPR012837	NrdG	11,645	4,135	630
IPR012838	PFL1_activating	9,425	2,995	160
IPR012839	Organic_radical_activase	16,374	5,941	1,127
IPR013483	MoaA	33,719	14,394	976

Fragment Option

UniProt designates a Sequence Status for each member: Complete if the encoding DNA sequence has both start and stop codons; Fragment if the start and/or stop codon is missing. Approximately 10% of the entries in UniProt are fragments.

Fragments: Check to exclude ← hits in the results. (default: off)

For the UniRef90 and UniRef50 databases, clusters are excluded if the cluster ID ("representative sequence") is a fragment.

UniProt IDs in UniRef90 and UniRef50 clusters with complete cluster IDs are removed from the clusters if they are fragments.

Filter by Taxonomy

From preselected conditions, the user can select "Bacteria, Archaea, Fungi", "Eukaryota, no Fungi", "Fungi", "Viruses", "Bacteria", "Eukaryota", or "Archaea" to restrict the retrieved sequences to these taxonomy groups.

"Bacteria, Archaea, Fungi", "Bacteria", "Archaea", and "Fungi" select organisms that may provide genome context (gene clusters/operons) useful for inferring functions.

The retrieved sequences also can be restricted to taxonomy categories within the Superkingdom, Kingdom, Phylum, Class, Order, Family, Genus, and Species ranks. Multiple conditions are combined to be a union of each other.

The sequences from the UniRef90 and UniRef50 databases are the UniRef90 and UniRef50 clusters for which the cluster ID ("representative sequence") matches the specified taxonomy categories. The UniProt members in these clusters that do not match the specified taxonomy categories are removed from the cluster.

Preselected conditions: -- select a preset to auto populate --

Add Taxonomy category

Protein Family Size Options

Family Domain Boundary Option

SSN Edge Calculation Option

Job name: (required) ←

E-mail address: ←

You will be notified by e-mail when your submission has been processed.

←

The SSN was finalized on the **SSN Finalization** tab of the **DATASET COMPLETED** page using 11 as the **Alignment Score Threshold** and 140 residues as the **Minimum** in the **Sequence Length Restriction** to remove truncated sequences, entering the **Network (SSN) name** (brown arrow), and clicking **“Create SSN”** (black arrow). We previously determined that the members of the anaerobic ribonucleotide reductase activating enzyme family have the shortest sequences (≥ 140 residues) [2].

DATASET COMPLETED

Submission Name: **IP91_RSS_UniRef90_NoFragments**

A minimum sequence similarity threshold that specifies the sequence pairs connected by edges is needed to generate the SSN. This threshold also determines the segregation of proteins into clusters. The threshold is applied to the edges in the SSN using the alignment score, an edge node attribute that is a measure of the similarity between sequence pairs.

Dataset Summary
Taxonomy Sunburst
Dataset Analysis
SSN Finalization

Alignment Score Threshold

This tab is used to specify the minimum "Alignment Score Threshold" (that is a measure of the minimum sequence similarity threshold) for drawing the edges that connect the proteins (nodes) in the SSN.

Alignment Score Threshold: ?

This value corresponds to the lower limit for which an edge will be present in the SSN. The alignment score is similar in magnitude to the negative base-10 logarithm of a BLAST e-value.

Sequence Length Restriction Options

Allows restriction of sequences in the generated SSN based on **sequence length**. ?

Minimum: (default: 0)

Maximum: (default: 50000)

▸ Filter by Taxonomy

▸ Neighborhood Connectivity

▸ Fragment Option

▸ Dev Site Options

Network name: This name will be displayed in Cytoscape.

You will be notified by e-mail when the SSN is ready for download.

Create SSN

The **DOWNLOAD NETWORK FILES** page did not provide the xgmml file for the full SSN or the representative node networks (348,446 nodes and 2,583,616,067 edges); the edge maximum for generating an SSN is 200,000,000.

SSN Overview
Network Files

Please cite your use of the EFI tools:

Rémi Zallot, Nils Oberg, and John A. Gerlt, **The EFI Web Resource for Genomic Enzymology Tools: Leveraging Protein, Genome, and Metagenome Databases to Discover Novel Enzymes and Metabolic Pathways**. *Biochemistry* 2019 58 (41), 4169-4182. <https://doi.org/10.1021/acs.biochem.9b00735>

The panels below provide files for full and representative node SSNs for download with the indicated numbers of nodes and edges. As an approximate guide, SSNs with ~2M edges can be opened with 16 GB RAM, ~5M edges can be opened with 32 GB RAM, ~10M edges can be opened with 64 GB RAM, ~20M edges can be opened with 128 GB RAM, ~40M edges can be opened with 256 GB RAM, and ~120M edges can be opened with 768 GB RAM.

Files may be transferred to the Genome Neighborhood Tool (GNT), the Color SSN utility, the Cluster Analysis utility, or the Neighborhood Connectivity utility.

Full Network ?

Each node in the network represents a single protein sequence.

The output file was too large (edges=2,583,616,067) to be generated by EST. Please use a repnode below or choose a different alignment score.

Representative Node Networks ?

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

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

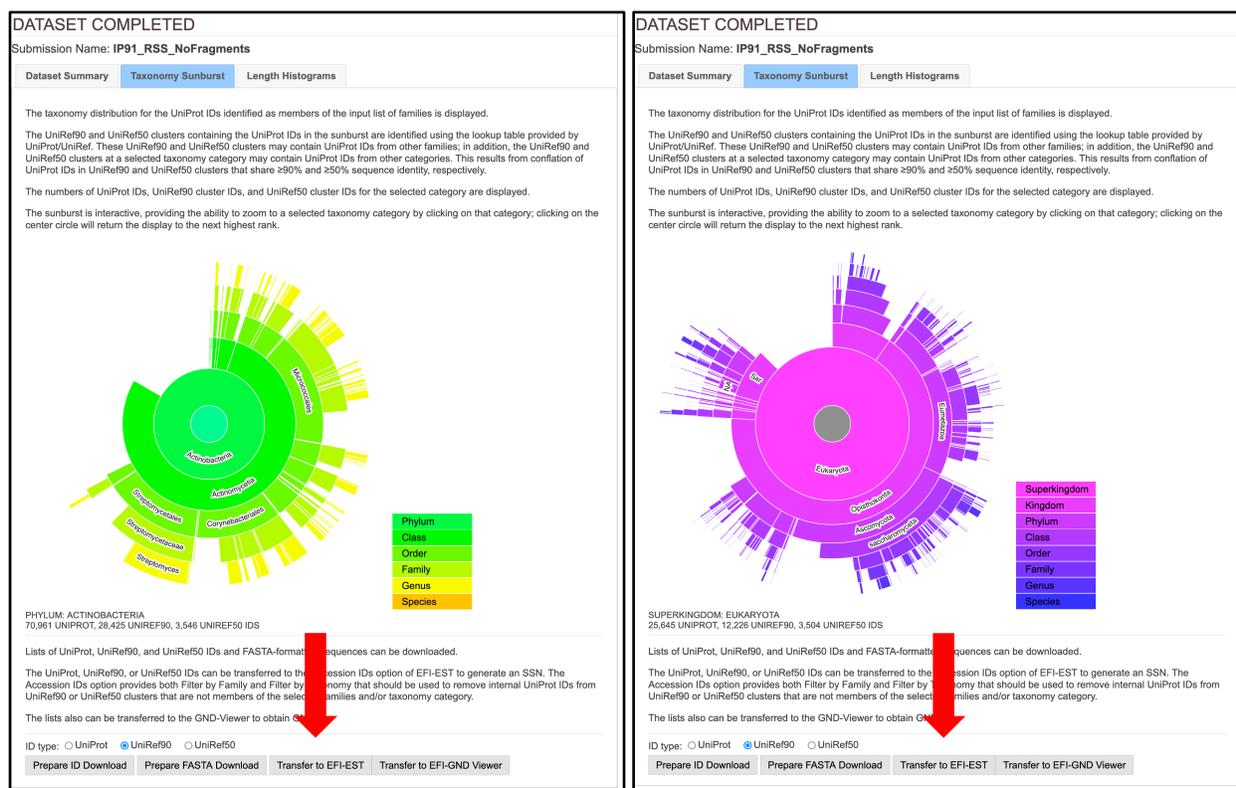
% ID	# Nodes	# Edges
100	The output file was too large (edges=2,583,616,067) to be generated by EST.	
95	The output file was too large (edges=2,569,142,064) to be generated by EST.	
90	The output file was too large (edges=2,526,601,854) to be generated by EST.	
85	The output file was too large (edges=2,165,619,179) to be generated by EST.	
80	The output file was too large (edges=1,846,597,151) to be generated by EST.	
75	The output file was too large (edges=1,577,543,078) to be generated by EST.	
70	The output file was too large (edges=1,333,628,047) to be generated by EST.	
65	The output file was too large (edges=1,099,873,487) to be generated by EST.	
60	The output file was too large (edges=886,414,047) to be generated by EST.	
55	The output file was too large (edges=695,711,057) to be generated by EST.	
50	The output file was too large (edges=527,455,035) to be generated by EST.	
45	The output file was too large (edges=389,443,099) to be generated by EST.	
40	The output file was too large (edges=267,527,451) to be generated by EST.	

Download Network Statistics as Table

New to Cytoscape?

Taxonomy Category-Specific UniRef90 SSNs: Taxonomy Tool Families Option, with Transfer of UniRef90 cluster IDs to the EFI-EST Accession IDs Option

The **Taxonomy Sunburst** for the entire RSS with complete sequences was used with the Transfer to EFI-EST feature to generate taxonomy category-specific UniRef90 SSNs that could be analyzed with Cytoscape. For superkingdom Bacteria, phylum Actinobacteria; superkingdom Bacteria, phylum Bacteroidetes; superkingdom Bacteria, phylum Firmicutes; superkingdom Bacteria, phylum Proteobacteria; and superkingdom Archaea, the indicated taxonomy categories were selected by clicking on the wedge (left panel). For **Preselected conditions** Fungi (a combination of four phyla within superkingdom Eukaryota), the Eukaryota taxonomy category was selected (right panel).



The UniRef90 cluster SSN for superkingdom Bacteria, phylum Proteobacteria is too large to be analyzed with Cytoscape (102,114 nodes and 250,587,566 edges). However, UniRef90 cluster SSNs were generated for Classes within the Proteobacteria that can be analyzed with Cytoscape: class Alphaproteobacteria, class Betaproteobacteria, class Gammaproteobacteria, class Deltaproteobacteria, and Class Epsilonproteobacteria.

For the single taxonomy categories (left panel), in the **EFI-EST Accession IDs Option** pages that opened, the **Fragment Option** was used to exclude fragments (blue arrow), **Filter by Family** was used with the list of 211 Pfam and InterPro families and/or domains (**Tutorial Table 1**) (green arrow), and **Filter by Taxonomy** was used to select the taxonomy category (magenta arrow). For Fungi, **Fungi** was selected from the **Preselected conditions** menu in **Filter by Taxonomy**; for Eukaryota, no Fungi, **Eukaryota, no Fungi** was selected from the **Preselected conditions** menu in **Filter by Taxonomy**. As described in the text, **Filter by Family** and **Filter by Taxonomy** are used to ensure that the UniRef90 cluster IDs and internal UniProt IDs match the desired taxonomy category and family. The **Job name** (orange arrow) and an **E-mail address** were entered (cyan arrow), and the job was started by clicking “**Create SSN**” (black arrow).

Previous Jobs Sequence BLAST Families FASTA **Accession IDs** SSN Utilities

Generate a SSN from a list of UniProt, UniRef, NCBI, or Genbank IDs.
 An all-by-all BLAST is performed to obtain the similarities between sequence pairs to calculate edge values to generate the SSN.

Use UniProt IDs Use UniRef50 or UniRef90 Cluster IDs

Input a list of UniRef50 or UniRef90 cluster accession IDs, or upload a text file.

Accession IDs:

Accession ID File:

RSS_NoFragments

Input accession IDs are: UniRef90 cluster IDs

Fragment Option

UniProt designates a Sequence Status for each member: Complete if the encoding DNA sequence has both start and stop codons; Fragment if the start and/or stop codon is missing. Approximately 10% of the entries in UniProt are fragments.

Fragments: Check to exclude UniProt results. (default: off)

For the UniRef90 and UniRef50 databases, clusters are excluded if the cluster ID ("representative sequence") is a fragment.

UniProt IDs in UniRef90 and UniRef50 clusters with complete cluster IDs are removed from the clusters if they are fragments.

Filter by Family

The input list of UniRef90 or UniRef50 cluster IDs should (must) be filtered with the same list of Pfam families, InterPro families, and/or Pfam clans used to generate the IDs, if:

The input list of UniRef90 or UniRef50 IDs is obtained from 1) the Color SSN or Cluster Analysis utility for a Families option (Option B) EPI-EST SSN, 2) the Families option of the Taxonomy Tool, or 3) the Accession IDs option of the Taxonomy Tool.

Input a list of Pfam families, InterPro families, and/or Pfam clans to restrict the UniProt and/or UniRef IDs in the SSN to these families.

Family(s): 5 PF06969 PF08497 PF12199 PF16881 PF19238 PF19288 PF19864

The input format is a single family or comma/space separated list of families. Families should be specified as PFxxxxx (five digits), IPRxxxxxx (six digits) or CLxxxx (four digits) for Pfam clans.

For input lists of UniRef90 and UniRef50 clusters, the cluster ID (representative sequence) is used to identify those that match the list of families and are included in the SSN. The UniProt members in these clusters that do not match the input families are removed from the cluster and are not included in the SSN node attributes.

Filter by Taxonomy

The input list of UniRef90 or UniRef50 cluster IDs should (must) be filtered with the same taxonomy categories used to generate the IDs, if:

The input list of UniRef90 or UniRef50 IDs is obtained from 1) the Color SSN or Cluster Analysis utility for a Families option (Option B) EPI-EST SSN, 2) the Families option of the Taxonomy Tool, or 3) the Accession IDs option of the Taxonomy Tool.

From preselected conditions, the user can select "Bacteria, Archaea, Fungi", "Eukaryota, no Fungi", "Fungi", "Viruses", "Bacteria", "Eukaryota", or "Archaea" to restrict the UniProt IDs in the sunburst to these taxonomy groups.

"Bacteria, Archaea, Fungi", "Bacteria", "Archaea", and "Fungi" select organisms that may provide genome context (gene clusters/operons) useful for inferring functions.

The UniProt IDs also can be restricted to taxonomy categories within the Superkingdom, Kingdom, Phylum, Class, Order, Family, Genus, and Species ranks. Multiple conditions are combined to be a union of each other.

Preselected conditions: -- select a preset to auto populate --

Phylum: Actinobacteria

Add Taxonomy category

Protein Family Addition Options

Family Domain Boundary Options

SSN Edge Calculation Option

Job name: acteria UniRef90_NoFragments_RSS_Actinobacteria (required)

E-mail address:

You will be notified by e-mail when your submission has been processed.

Submit Analysis

Previous Jobs Sequence BLAST Families FASTA **Accession IDs** SSN Utilities

Generate a SSN from a list of UniProt, UniRef, NCBI, or Genbank IDs.
 An all-by-all BLAST is performed to obtain the similarities between sequence pairs to calculate edge values to generate the SSN.

Use UniProt IDs Use UniRef50 or UniRef90 Cluster IDs

Input a list of UniRef50 or UniRef90 cluster accession IDs, or upload a text file.

Accession IDs:

Accession ID File:

IP91_RSS_NoFragments

Input accession IDs are: UniRef90 cluster IDs

Fragment Option

UniProt designates a Sequence Status for each member: Complete if the encoding DNA sequence has both start and stop codons; Fragment if the start and/or stop codon is missing. Approximately 10% of the entries in UniProt are fragments.

Fragments: Check to exclude UniProt-defined results. (default: off)

For the UniRef90 and UniRef50 databases, clusters are excluded if the cluster ID ("representative sequence") is a fragment.

UniProt IDs in UniRef90 and UniRef50 clusters with complete cluster IDs are removed from the clusters if they are fragments.

Filter by Family

The input list of UniRef90 or UniRef50 cluster IDs should (must) be filtered with the same list of Pfam families, InterPro families, and/or Pfam clans used to generate the IDs, if:

The input list of UniRef90 or UniRef50 IDs is obtained from 1) the Color SSN or Cluster Analysis utility for a Families option (Option B) EPI-EST SSN, 2) the Families option of the Taxonomy Tool, or 3) the Accession IDs option of the Taxonomy Tool.

Input a list of Pfam families, InterPro families, and/or Pfam clans to restrict the UniProt and/or UniRef IDs in the SSN to these families.

Family(s): 5 PF06969 PF08497 PF12345 PF12345 PF19238 PF19288 PF19864

The input format is a single family or comma/space separated list of families. Families should be specified as PFxxxxx (five digits), IPRxxxxxx (six digits) or CLxxxx (four digits) for Pfam clans.

For input lists of UniRef90 and UniRef50 clusters, the cluster ID (representative sequence) is used to identify those that match the list of families and are included in the SSN. The UniProt members in these clusters that do not match the input families are removed from the cluster and are not included in the SSN node attributes.

Filter by Taxonomy

The input list of UniRef90 or UniRef50 cluster IDs should (must) be filtered with the same taxonomy categories used to generate the IDs, if:

The input list of UniRef90 or UniRef50 IDs is obtained from 1) the Color SSN or Cluster Analysis utility for a Families option (Option B) EPI-EST SSN, 2) the Families option of the Taxonomy Tool, or 3) the Accession IDs option of the Taxonomy Tool.

From preselected conditions, the user can select "Bacteria, Archaea, Fungi", "Eukaryota, no Fungi", "Fungi", "Viruses", "Bacteria", "Eukaryota", or "Archaea" to restrict the UniProt IDs in the sunburst to these taxonomy groups.

"Bacteria, Archaea, Fungi", "Bacteria", "Archaea", and "Fungi" select organisms that may provide genome context (gene clusters/operons) useful for inferring functions.

The UniProt IDs also can be restricted to taxonomy categories within the Superkingdom, Kingdom, Phylum, Class, Order, Family, Genus, and Species ranks. Multiple conditions are combined to be a union of each other.

Preselected conditions: Fungi

Phylum: Ascomycota

Phylum: Basidiomycota

Phylum: Fungi incertae sedis

Phylum: unclassified fungi

Reset

Protein Family Addition Options

Family Domain Boundary Options

SSN Edge Calculation Option

Job name: tents Eukaryota UniRef90_NoFragments_RSS_Fungi (required)

E-mail address:

You will be notified by e-mail when your submission has been processed.

Submit Analysis

The SSNs were finalized on the **SSN Finalization** tab of the **DATASET COMPLETED** page using 11 as the **Alignment Score Threshold** and 140 residues as the **Minimum** in the **Sequence Length Restriction** to remove truncated sequences (orange arrow), entering the **Network (SSN) name** (brown arrow), and clicking “**Create SSN**” (black arrow). We previously determined that the members of the anaerobic ribonucleotide reductase activating enzyme family have the shortest sequences (≥ 140 residues) [2].

DATASET COMPLETED

Submission Name: **IP91_RSS_NoFragments_Actinobacteria_UniRef90_NoFragments_RSS_Actinobacteria**

A minimum sequence similarity threshold that specifies the sequence pairs connected by edges is needed to generate the SSN. This threshold also determines the segregation of proteins into clusters. The threshold is applied to the edges in the SSN using the alignment score, an edge node attribute that is a measure of the similarity between sequence pairs.

Dataset Summary
Taxonomy Sunburst
Dataset Analysis
SSN Finalization

This tab is used to specify the minimum "Alignment Score Threshold" (that is a measure of the minimum sequence similarity threshold) for drawing the edges that connect the proteins (nodes) in the SSN.

Alignment Score Threshold: ?

This value corresponds to the lower limit for which an edge will be present in the SSN. The alignment score is similar in magnitude to the negative base-10 logarithm of a BLAST e-value.

▾ **Sequence Length Restriction Options**

Allows restriction of sequences in the generated SSN based on their length. ?

Minimum: (default: 0)

Maximum: (default: 50000)

▸ Filter by Taxonomy

▸ Neighborhood Connectivity

▸ Fragment Option

Network name: This name will be used in Cytoscape.

You will be notified by e-mail when the SSN is ready for download.

Create SSN

The **Network Files** tab of the **DOWNLOAD NETWORK FILES** page provided the xgmml file for the **Full (SSN) Network** (red arrow; all UniProt nodes and edges) as well as the xgmml files for **Representative Node Networks** that conflate the UniProt nodes based on percent identity (blue arrow). The xgmml files for the full SSN was downloaded, opened with Cytoscape 3.9.1, and displayed with the Prefuse Force Directed layout. The nodes were colored according to the subgroups defined by the Structure-Function Linkage Database (SFLD) [2, 3].

DOWNLOAD NETWORK FILES

Submission Name: IP91_RSS_NoFragments_Actinobacteria_UniRef90_NoFragments_RSS_Actinobacteria
 Network Name: IP91_RSS_NoFragments_Actinobacteria_UniRef90_NoFragments_RSS_Actinobacteria_Minlen140_AS11

SSN Overview **Network Files**

Please cite your use of the EFI tools:

Rémi Zallot, Nils Oberg, and John A. Gerit, **The EFI Web Resource for Genomic Enzymology Tools: Leveraging Protein, Genome, and Metagenome Databases to Discover Novel Enzymes and Metabolic Pathways.** *Biochemistry* 2019 58 (41), 4169-4182. <https://doi.org/10.1021/acs.biochem.9b00735>

The panels below provide files for full and representative node SSNs for download with the indicated numbers of nodes and edges. As an approximate guide, SSNs with ~2M edges can be opened with 16 GB RAM, ~5M edges can be opened with 32 GB RAM, ~10M edges can be opened with 64 GB RAM, ~20M edges can be opened with 128 GB RAM, ~40M edges can be opened with 256 GB RAM, and ~120M edges can be opened with 768 GB RAM.

Files may be transferred to the Genome Neighborhood Tool (GNT), the Color SSN utility, the Cluster Analysis utility, or the Neighborhood Connectivity utility.

Full Network ← ?

Each node in the network represents a single protein sequence.

	# Nodes	# Edges	
Download ZIP	27,953	20,427,812	Transfer To: ▼

Representative Node Networks ← ?

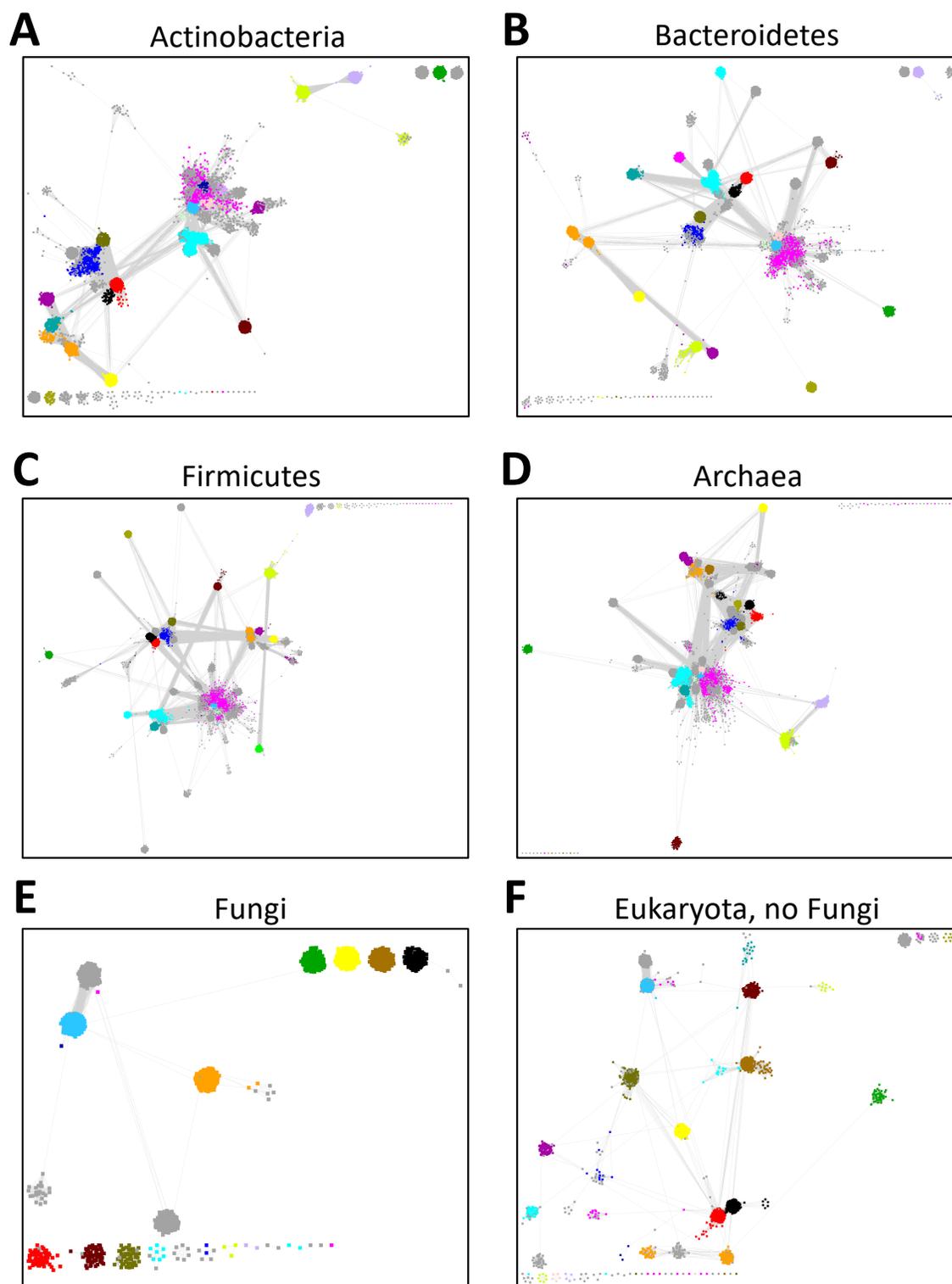
In representative node (RepNode) networks, each node in the network represents a collection of proteins grouped according to percent identity. For example, for a 75% identity RepNode network, all connected sequences that share 75% or more identity are grouped into a single node (meta node). Sequences are collapsed together to reduce the overall number of nodes, making for less complicated networks easier to load in Cytoscape.

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

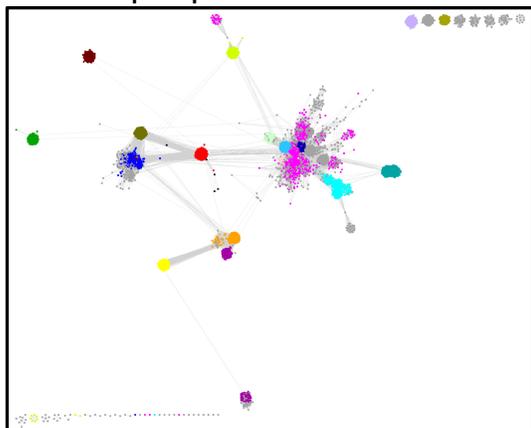
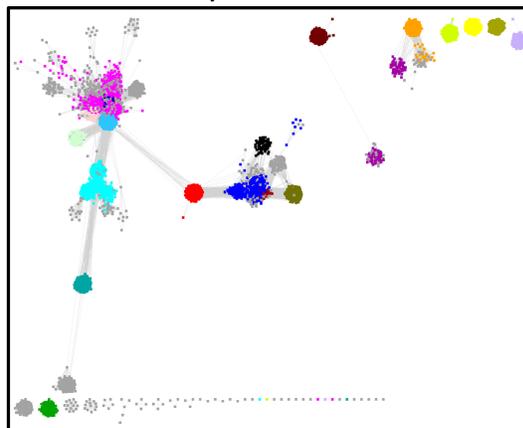
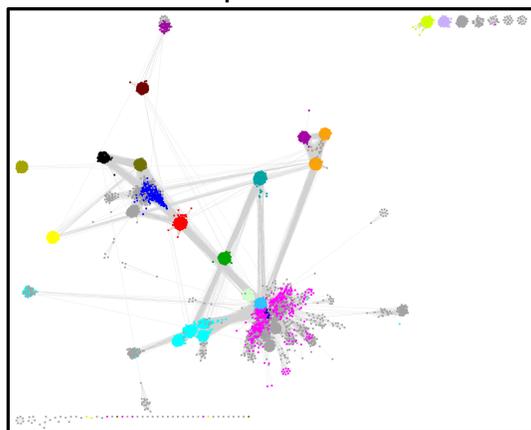
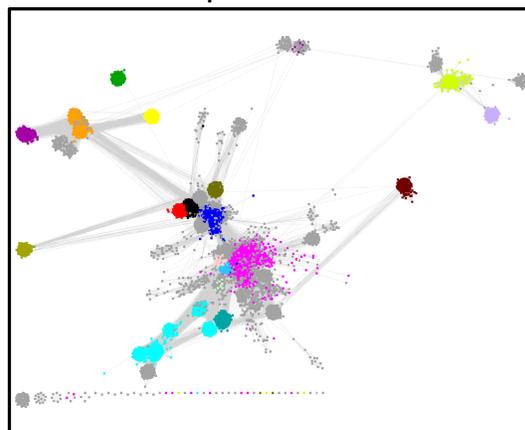
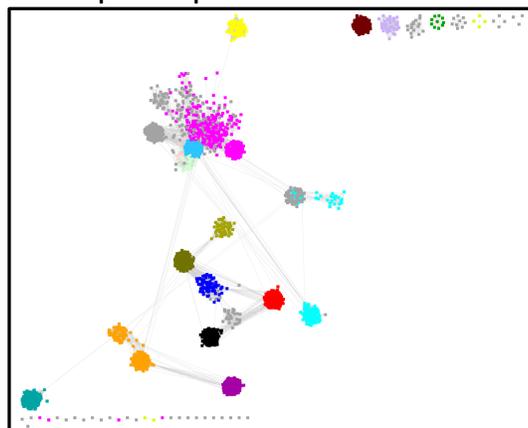
	% ID	# Nodes	# Edges	
Download ZIP	100	27,953	20,427,812	Transfer To: ▼
Download ZIP	95	27,810	20,211,143	Transfer To: ▼
Download ZIP	90	27,432	19,597,797	Transfer To: ▼
Download ZIP	85	24,941	15,869,268	Transfer To: ▼
Download ZIP	80	22,566	12,706,893	Transfer To: ▼
Download ZIP	75	20,347	9,915,420	Transfer To: ▼
Download ZIP	70	18,190	7,522,808	Transfer To: ▼
Download ZIP	65	16,043	5,489,593	Transfer To: ▼
Download ZIP	60	14,094	3,955,693	Transfer To: ▼
Download ZIP	55	12,480	2,924,263	Transfer To: ▼
Download ZIP	50	11,215	2,264,911	Transfer To: ▼
Download ZIP	45	10,258	1,841,807	Transfer To: ▼
Download ZIP	40	9,409	1,496,401	Transfer To: ▼

Download Network Statistics as Table

New to Cytoscape?



Taxonomy Category-Filtered UniRef 90 Cluster SSNs for the RS Superfamily. The SSNs were generated as described in the text. The SSNs were generated using a minimum length of 140 residues and an alignment score threshold of 11. The nodes are colored according to the subgroups defined by the Structure-Function Linkage Database (SFLD) [2, 3]. **Panel A**, Superkingdom Bacteria, phylum Actinobacteria; the SSN contains 27,953 nodes and 20,427,984 edges. **Panel B**, Superkingdom Bacteria, phylum Bacteroidetes; the SSN contains 27,028 nodes and 21,948,018 edges. **Panel C**, Superkingdom Bacteria, phylum Firmicutes; the SSN contains 52,451 nodes and 61,254,499 edges. **Panel D**, Superkingdom Archaea; the SSN contains 36,996 nodes and 27,905,053 edges. **Panel E**, Superkingdom Eukaryota, Fungi only; the SSN contains 3,341 nodes and 715,572 edges. **Panel F**, Superkingdom Eukaryota, no Fungi; the SSN contains 8,2661 nodes and 3,859,619 edges.

A Alphaproteobacteria**B** Betaproteobacteria**C** Gammaproteobacteria**D** Deltaproteobacteria**E** Epsilonproteobacteria

Taxonomy Category-Filtered UniRef 90 Cluster SSNs for Taxonomy Classes in the Proteobacteria Phylum in the Radical SAM Superfamily. The SSNs were generated using the **Families option of the Taxonomy Tool with transfer of UniRef90 cluster IDs to Option D** pipeline described in the text. The SSNs were generated using a minimum length of 140 residues and an alignment score threshold of 11. The nodes are colored according to the subgroups defined by the Structure-Function Linkage Database (SFLD) [2, 3]. **Panel A**, Class Alphaproteobacteria; the SSN contains 27,868 nodes and 29,935,838 edges. **Panel B**, Class Betaproteobacteria; the SSN contains 11,936 nodes and 4,364,852 edges. **Panel C**, Class Gammaproteobacteria; the SSN contains 28,350 nodes and 24,151,979 edges. **Panel D**, Class Deltaproteobacteria; the SSN contains 26,875 nodes and 15,107,300 edges. **Panel E**, Class Epsilonproteobacteria; the SSN contains 3,830 nodes and 513,294 edges.

Taxonomy Category-Specific UniRef90 SSNs: EFI-EST Families Option, Filter by Taxonomy in the Analysis Step

The SSN Finalization tab of the **DATASET COMPLETED** page for the UniRef90 cluster SSN for the complete entries was used to generate the category-filtered SSNs described in the **Taxonomy Tool Families Option, with transfer of UniRef90 cluster IDs to the EFI-EST Accession IDs Option** section. The SSNs were finalized using 11 as the **Alignment Score Threshold** and 140 residues as the **Minimum** in the **Sequence Length Restriction** to remove truncated sequences. As described previously, for the single taxonomy categories, **Filter by Taxonomy** was used to select the single taxonomy categories. For Fungi, **Fungi** was selected from the **Preselected conditions** menu. For Eukaryota, no Fungi, **Eukaryota, no Fungi** was selected from the **Preselected conditions** menu.

Dataset Summary Taxonomy Sunburst Dataset Analysis **SSN Finalization**

Alignment Score Threshold

This tab is used to specify the minimum "Alignment Score Threshold" (that is a measure of the minimum sequence similarity threshold) for drawing the edges that connect the proteins (nodes) in the SSN.

Alignment Score Threshold:

This value corresponds to the lower limit for which an edge will be present in the SSN. The alignment score is similar in magnitude to the negative base-10 logarithm of a BLAST e-value.

Sequence Length Restriction Options

Allows restriction of sequences in the generated SSN based on their length.

Minimum: (default: 0)

Maximum: (default: 50000)

Filter by Taxonomy

From preselected conditions, the user can select "Bacteria, Archaea, Fungi", "Eukaryota, no Fungi", "Fungi", "Viruses", "Bacteria", "Eukaryota", or "Archaea" to restrict the SSN nodes to these taxonomy groups.

"Bacteria, Archaea, Fungi", "Bacteria", "Archaea", and "Fungi" select organisms that may provide genome context (gene clusters/operons) useful for inferring functions.

The SSN nodes also can be restricted to taxonomy categories within the Superkingdom, Kingdom, Phylum, Class, Order, Family, Genus, and Species ranks. Multiple conditions are combined to be a union of each other.

The SSN nodes from the UniRef90 and UniRef50 databases are the UniRef90 and UniRef50 clusters for which the cluster ID ("representative sequence") matches the specified taxonomy categories. The UniProt members in these nodes that do not match the specified taxonomy categories are removed from the nodes.

Preselected conditions: -- select a preset to auto populate --

Phylum:

Add taxonomic condition

Neighborhood Connectivity

Fragment Option

Dev Site Options

Network name: `'ef90_NoFragments_Actinobacteria_Minlen140_AS11'` This name will be displayed in Cytoscape.

You will be notified by e-mail when the SSN is ready for download.

Dataset Summary Taxonomy Sunburst Dataset Analysis **SSN Finalization**

Alignment Score Threshold

This tab is used to specify the minimum "Alignment Score Threshold" (that is a measure of the minimum sequence similarity threshold) for drawing the edges that connect the proteins (nodes) in the SSN.

Alignment Score Threshold:

This value corresponds to the lower limit for which an edge will be present in the SSN. The alignment score is similar in magnitude to the negative base-10 logarithm of a BLAST e-value.

Sequence Length Restriction Options

Allows restriction of sequences in the generated SSN based on their length.

Minimum: (default: 0)

Maximum: (default: 50000)

Filter by Taxonomy

From preselected conditions, the user can select "Bacteria, Archaea, Fungi", "Eukaryota, no Fungi", "Fungi", "Viruses", "Bacteria", "Eukaryota", or "Archaea" to restrict the SSN nodes to these taxonomy groups.

"Bacteria, Archaea, Fungi", "Bacteria", "Archaea", and "Fungi" select organisms that may provide genome context (gene clusters/operons) useful for inferring functions.

The SSN nodes also can be restricted to taxonomy categories within the Superkingdom, Kingdom, Phylum, Class, Order, Family, Genus, and Species ranks. Multiple conditions are combined to be a union of each other.

The SSN nodes from the UniRef90 and UniRef50 databases are the UniRef90 and UniRef50 clusters for which the cluster ID ("representative sequence") matches the specified taxonomy categories. The UniProt members in these nodes that do not match the specified taxonomy categories are removed from the nodes.

Preselected conditions: Fungi

Phylum:

Phylum:

Phylum:

Phylum:

Neighborhood Connectivity

Fragment Option

Dev Site Options

Network name: `'1_RSS_UniRef90_NoFragments_Fungi_Minlen140_A'` This name will be displayed in Cytoscape.

You will be notified by e-mail when the SSN is ready for download.

The xgmml files for the full UniRef90 cluster SSNs (all UniRef90 cluster nodes and edges with alignment scores ≥ 11) as well as representative node networks that conflate UniRef90 clusters nodes based on percent identity were available for download on the **DOWNLOAD NETWORK FILES** page. The xgmml files for the full SSNs were downloaded and opened with Cytoscape; the nodes were colored according to the subgroups defined by the Structure-Function Linkage Database (SFLD) [2, 3].

DOWNLOAD NETWORK FILES

Submission Name: **IP91_RSS_UniRef90_NoFragments**
 Network Name: **IP91_RSS_UniRef90_NoFragments_Actinobacteria_Minlen140_AS11**

SSN Overview **Network Files**

Please cite your use of the EFI tools:

Rémi Zallot, Nils Oberg, and John A. Gerit, **The EFI Web Resource for Genomic Enzymology Tools: Leveraging Protein, Genome, and Metagenome Databases to Discover Novel Enzymes and Metabolic Pathways**. *Biochemistry* 2019 58 (41), 4169-4182. <https://doi.org/10.1021/acs.biochem.9b00735>

The panels below provide files for full and representative node SSNs for download with the indicated numbers of nodes and edges. As an approximate guide, SSNs with ~2M edges can be opened with 16 GB RAM, ~5M edges can be opened with 32 GB RAM, ~10M edges can be opened with 64 GB RAM, ~20M edges can be opened with 128 GB RAM, ~40M edges can be opened with 256 GB RAM, and ~120M edges can be opened with 768 GB RAM.

Files may be transferred to the Genome Neighborhood Tool (GNT), the Color SSN utility, the Cluster Analysis utility, or the Neighborhood Connectivity utility.

Full Network ?

Each node in the network represents a single protein sequence.

	# Nodes	# Edges	
Download ZIP	27,953	20,397,534	Transfer To: ▾

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

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

	% ID	# Nodes	# Edges	
Download ZIP	100	27,953	20,397,534	Transfer To: ▾
Download ZIP	95	27,810	20,180,999	Transfer To: ▾
Download ZIP	90	27,432	19,568,000	Transfer To: ▾
Download ZIP	85	24,941	15,842,539	Transfer To: ▾
Download ZIP	80	22,566	12,682,765	Transfer To: ▾
Download ZIP	75	20,347	9,893,597	Transfer To: ▾
Download ZIP	70	18,190	7,503,796	Transfer To: ▾
Download ZIP	65	16,043	5,472,899	Transfer To: ▾
Download ZIP	60	14,094	3,940,841	Transfer To: ▾
Download ZIP	55	12,480	2,911,030	Transfer To: ▾
Download ZIP	50	11,215	2,253,155	Transfer To: ▾
Download ZIP	45	10,258	1,831,127	Transfer To: ▾
Download ZIP	40	9,409	1,486,723	Transfer To: ▾

Download Network Statistics as Table

New to Cytoscape?

Taxonomy Category-Specific UniRef90 SSNs: EFI-EST Families Option, Filter by Taxonomy in the Generate Step

The same taxonomy category-filtered UniRef90 SSNs described in the previous sections were generated in separate jobs using the **EFI-EST Family Option** by specifying the list of 211 Pfam and InterPro families and/or domains (**Tutorial Table 1**) (red arrow) and UniRef50 cluster IDs (blue arrow), selecting **Fragment Option** to exclude fragments (green arrow), and selecting the taxonomy category (magenta arrow). The **Job name** (orange arrow) and an **E-mail address** (cyan arrow) were entered; the job was started by clicking “**Submit analysis**” (black arrow).

Generate a SSN for a protein family.
The members of the input Pfam families, InterPro families, and/or Pfam clans are selected from the UniProt, UniRef90, or UniRef50 database.

Pfam and/or InterPro Families and/or domains: **45784 PF04055 PF06969 PF08007 PF12345 PF13186 PF16199 PF16881 PF19238 PF19288 PF19864**

Use UniRef90 cluster ID (representative sequence) or UniRef50 cluster IDs (UniProt is default).

Family	Family Name	Full Size	UniRef90 Size	UniRef50 Size
IPR000385	MoaA_NiFB_PqqE_Fe-S-bd_CS	49,241	23,160	4,777
IPR001989	Radical_activat_CS	26,935	9,907	1,836
IPR002684	Biotin_synth/BioAB	27,640	9,880	1,004
IPR003698	Lipoyl_synth	39,047	13,924	1,318
IPR003739	Lys_aminomutase/Glu_NH3_mut	20,775	10,372	1,278
IPR004383	rRNA_Isu_MTrfase_RimN/Cfr	39,944	15,429	1,455
IPR004558	Coprogen_oxidase_HemN	16,796	6,746	513
IPR004559	HemN-like	38,255	17,990	2,765
IPR005839	Methylthioesterase	87,716	37,295	4,127
IPR005840	Ribosomal_S12_MeSTrfase_RimO	28,658	11,857	2,029
IPR005909	RaSEA	2,035	1,022	229
IPR005911	YhcC-like	11,293	4,505	502
IPR005980	Nase_CF_NiFB	2,647	1,104	82
IPR006463	MiaB_methylolase	35,615	13,649	723
IPR006466	MiaB-like_B	4,407	2,226	506
IPR006467	MiaB-like_C	17,077	8,216	1,089
IPR006638	Elp3/MiaA/NiFB-like_rSAM	446,282	212,389	36,535
IPR007197	rSAM	722,535	355,669	70,723
IPR010505	Mob_synth_C	38,361	16,635	1,682
IPR010722	BATS_dom	39,852	14,953	1,222
IPR010723	HemN_C	39,495	17,002	2,804
IPR011101	DUF5131	7,313	4,801	1,324
IPR011843	PQQ_synth_PqqE_bac	5,549	1,835	59
IPR012726	ThiH	6,704	2,297	125
IPR012837	NrdG	11,645	4,135	630
IPR012838	PFL1_activating	9,425	2,995	180
IPR012839	Organic_radical_activase	16,374	5,941	1,127
IPR013483	MoaA	33,719	14,394	976
IPR013704	UPF0313_N	12,479	4,760	404
IPR013848	Methylthioesterase_N	91,463	39,689	5,468

Fragment Option
UniProt designates a Sequence Status for each member: Complete if the encoding DNA sequence has both start and stop codons; Fragment if the start and/or stop codon is missing. Approximately 10% of the entries in UniProt are fragments.
Fragments: Check to exclude fragments in the results. (default: off)
For the UniRef90 and UniRef50 databases, clusters are excluded if the cluster ID ("representative sequence") is a fragment.
UniProt IDs in UniRef90 and UniRef50 clusters with complete cluster IDs are removed from the clusters if they are fragments.

Filter by Taxonomy
From preselected conditions, the user can select "Bacteria, Archaea, Fungi", "Eukaryota, no Fungi", "Fungi", "Viruses", "Bacteria", "Eukaryota", or "Archaea" to restrict the retrieved sequences to these taxonomy groups.
"Bacteria, Archaea, Fungi", "Bacteria", "Archaea", and "Fungi" select organisms that may provide genome context (gene clusters/operons) useful for inferring functions.
The retrieved sequences also can be restricted to taxonomy categories within the Superkingdom, Kingdom, Phylum, Class, Order, Family, Genus, and Species ranks. Multiple conditions are combined to be a union of each other.
The sequences from the UniRef90 and UniRef50 databases are the UniRef90 and UniRef50 clusters for which the cluster ID ("representative sequence") matches the specified taxonomy categories. The UniProt members in these clusters that do not match the specified taxonomy categories are removed from the cluster.
Preselected conditions: -- select a preset to auto populate --
Phylum: Actinobacteria
Add Taxonomy category

Protein Family Size Options
Family Domain Boundary Option
SSN Edge Calculation Option

Job name: **IP91_RSS_UniRef90_NoFragments_Actinobacteria** (required)
E-mail address:
You will be notified by e-mail when your submission has been processed.
Submit Analysis

The SSNs were finalized (**SSN Finalization** tab on the **DATASET COMPLETED** pages) using 11 as the **Alignment Score Threshold** (orange arrow) and 140 residues as the **Minimum** in the **Sequence Length Restriction** to remove truncated sequences (cyan arrow), entering the **Network (SSN) name** (brown arrow), and clicking “**Create SSN**” (black arrow). We previously determined that the members of the anaerobic ribonucleotide reductase activating enzyme family have the shortest sequences (≥ 140 residues) [2].

DATASET COMPLETED

Submission Name: **IP91_RSS_UniRef90_NoFragments_Actinobacteria**

A minimum sequence similarity threshold that specifies the sequence pairs connected by edges is needed to generate the SSN. This threshold also determines the segregation of proteins into clusters. The threshold is applied to the edges in the SSN using the alignment score, an edge node attribute that is a measure of the similarity between sequence pairs.

Dataset Summary

Taxonomy Sunburst

Dataset Analysis

SSN Finalization

This tab is used to specify the minimum "Alignment Score Threshold" (that is a measure of the minimum sequence similarity threshold) for drawing the edges that connect the proteins (nodes) in the SSN.

Alignment Score Threshold: ?

This value corresponds to the lower limit for which an edge will be present in the SSN. The alignment score is similar in magnitude to the negative base-10 logarithm of a BLAST e-value.

▾ Sequence Length Restriction Options

Allows restriction of sequences in the generated SSN based on their length. ?

Minimum: (default: 0)

Maximum: (default: 50000)

▸ Filter by Taxonomy

▸ Neighborhood Connectivity

▸ Fragment Option

Network name: This name will be used in Cytoscape.

You will be notified by e-mail when the SSN is ready for download.

Create SSN

The **Network Files** tab of the **DOWNLOAD NETWORK FILES** pages provided the xgmml file for the **Full (SSN) Network** (red arrow; all UniProt nodes and edges) as well as the xgmml files for **Representative Node Networks** that conflate the UniProt nodes based on percent identity (blue arrow). The xgmml files for the full SSNs were downloaded, opened with Cytoscape 3.9.1, and displayed with the Prefuse Force Directed layout; the nodes were colored according to the Structure-Function Linkage Database subgroups [2, 3].

DOWNLOAD NETWORK FILES

Submission Name: IP91_RSS_UniRef90_NoFragments_Actinobacteria
 Network Name: IP91_RSS_UniRef90_NoFragments_Actinobacteria_Minlen140_AS11

SSN Overview **Network Files**

Please cite your use of the EFI tools:

Rémi Zallot, Nils Oberg, and John A. Gerit, **The EFI Web Resource for Genomic Enzymology Tools: Leveraging Protein, Genome, and Metagenome Databases to Discover Novel Enzymes and Metabolic Pathways**. *Biochemistry* 2019 58 (41), 4169-4182. <https://doi.org/10.1021/acs.biochem.9b00735>

The panels below provide files for full and representative node SSNs for download with the indicated numbers of nodes and edges. As an approximate guide, SSNs with ~2M edges can be opened with 16 GB RAM, ~5M edges can be opened with 32 GB RAM, ~10M edges can be opened with 64 GB RAM, ~20M edges can be opened with 128 GB RAM, ~40M edges can be opened with 256 GB RAM, and ~120M edges can be opened with 768 GB RAM.

Files may be transferred to the Genome Neighborhood Tool (GNT), the Color SSN utility, the Cluster Analysis utility, or the Neighborhood Connectivity utility.

Full Network ← ?

Each node in the network represents a single protein sequence.

	# Nodes	# Edges	
Download ZIP	27,953	20,427,984	Transfer To: ▼

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

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

	% ID	# Nodes	# Edges	
Download ZIP	100	27,953	20,427,984	Transfer To: ▼
Download ZIP	95	27,810	20,211,304	Transfer To: ▼
Download ZIP	90	27,432	19,597,945	Transfer To: ▼
Download ZIP	85	24,941	15,869,380	Transfer To: ▼
Download ZIP	80	22,566	12,707,063	Transfer To: ▼
Download ZIP	75	20,347	9,915,581	Transfer To: ▼
Download ZIP	70	18,190	7,522,943	Transfer To: ▼
Download ZIP	65	16,043	5,489,724	Transfer To: ▼
Download ZIP	60	14,094	3,955,833	Transfer To: ▼
Download ZIP	55	12,480	2,924,366	Transfer To: ▼
Download ZIP	50	11,215	2,265,016	Transfer To: ▼
Download ZIP	45	10,258	1,841,906	Transfer To: ▼
Download ZIP	40	9,409	1,496,503	Transfer To: ▼

Download Network Statistics as Table

New to Cytoscape?

Tutorial Table 1. Pfam and IntePro Families Used to Identify Member of the RSS.

ID	Short Name	ID	Short Name
IPR000385	MoaA_NifB_PqqE_Fe-S-bd_CS	IPR026423	rSAM_cobopep
IPR001989	Radical_activat_CS	IPR026426	rSAM_FibroRumin
IPR002684	Biotin_synth/BioAB	IPR026429	MIA_synthase
IPR003698	Lipoyl_synth	IPR026447	B12_SAM-Ta0216
IPR003739	Lys_aminomutase/Glu_NH3_mut	IPR026482	rSAM_nif11_3
IPR004383	rRNA_Isu_MTrfase_RlmN/Cfr	IPR027492	RNA_MTrfase_RlmN
IPR004558	Coprogen_oxidase_HemN	IPR027526	Lipoyl_synth_chlpt
IPR004559	HemW-like	IPR027527	Lipoyl_synth_mt
IPR005839	Methylthiotransferase	IPR027559	B12_rSAM_oligo
IPR005840	Ribosomal_S12_MeSTrfase_RimO	IPR027564	HpnR_B12_rSAM
IPR005909	RaSEA	IPR027570	GeoRSP_rSAM
IPR005911	YhcC-like	IPR027583	rSAM_ACGX
IPR005980	Nase_CF_NifB	IPR027586	rSAM_metal_mat
IPR006463	MiaB_methiolase	IPR027596	AmmeMemoSam_rS
IPR006466	MiaB-like_B	IPR027604	W_rSAM_matur
IPR006467	MiaB-like_C	IPR027608	Spiro_SPASM
IPR006638	Elp3/MiaB/NifB	IPR027609	rSAM_QueE_Proteobac
IPR007197	rSAM	IPR027621	rSAM_QueE_gams
IPR010505	Mob_synth_C	IPR027622	rSAM_Clo7bot
IPR010722	BATS_dom	IPR027626	Pseudo_SAM_Halo
IPR010723	HemN_C	IPR027633	rSAM_NirJ2
IPR011101	DUF5131	IPR030801	Glu_2_3_NH3_mut
IPR011843	PQQ_synth_PqqE_bac	IPR030837	B12_rSAM_cofa1
IPR012726	ThiH	IPR030894	Ahb_Proteobacteria
IPR012837	NrdG	IPR030896	rSAM_AhbD_hemeb
IPR012838	PFL1_activating	IPR030905	CutC_activ_rSAM
IPR012839	Organic_radical_activase	IPR030915	rSAM_SkfB
IPR013483	MoaA	IPR030933	Non_iron_rSAM
IPR013704	UPF0313_N	IPR030950	rSAM_PoyD
IPR013848	Methylthiotransferase_N	IPR030969	B12_rSAM_trp_MT
IPR013917	tRNA_wybutosine-synth	IPR030977	QueE_Cx14CxxC
IPR014191	Anaer_RNR_activator	IPR030989	rSAM_XyeB
IPR016431	Pyrv-formate_lyase-activ_prd	IPR031003	BcpD_PhpK_rSAM
IPR016771	Fe-S_OxRdtase_rSAM_TM0948_prd	IPR031004	rSAM_YfkAB
IPR016779	rSAM_MSMEG0568	IPR031010	rSAM_mob_pairA
IPR016863	DesII	IPR031012	rSAM_mob_pairB
IPR017200	PqqE-like	IPR031014	rSAM_BlsE

IPR017672	MA_4551-like	IPR031015	Arg_2_3_am_muta
IPR017742	Deazaguanine_synth	IPR031019	rSAM_vs_C_rich
IPR017833	Hopanoid_synth-assoc_rSAM_HpnH	IPR031691	LIAS_N
IPR017834	Hopanoid_synth-assoc_rSAM_HpnJ	IPR032432	Radical_SAM_C
IPR019939	CofG_family	IPR033971	Avilamycin_epimerase
IPR019940	CofH_family	IPR033974	Glycerol_dehydratase_activase
IPR020050	FO_synthase_su2	IPR033975	ThnP-like
IPR020612	Methylthiotransferase_CS	IPR033976	GntE-like
IPR022431	Cyclic_DHFL_synthase_mqnC	IPR034165	NifB_C
IPR022432	MqnE	IPR034386	BtrN-like
IPR022447	Lys_aminomutase-rel	IPR034391	Cmo-like_SPASM_containing
IPR022459	Lysine_aminomutase	IPR034405	F420
IPR022462	EpmB	IPR034422	HydE/PyIB-like
IPR022881	rRNA_Isu_MeTfrase_Cfr	IPR034428	ThiH/NoCL/HydG-like
IPR022946	UPF0313	IPR034436	NocN/NosN-like
IPR023404	rSAM_horseshoe	IPR034438	4-hPhe_decarboxylase_activase
IPR023805	Uncharacterised_Spl-rel	IPR034457	Organic_radical-activating
IPR023807	Peptide_mod_rSAM	IPR034462	Benzylsuc_synthase_activase
IPR023819	Pep-mod_rSAM_AF0577	IPR034465	Pyruvate_for-lyase_activase
IPR023820	rSAM_GDL-assoc	IPR034466	Methyltransferase_Class_B
IPR023821	rSAM_TatD-assoc	IPR034471	7_8-dihydro-6-hydroxymethylpte
IPR023822	rSAM_TatD-assoc_bac	IPR034474	Methyltransferase_Class_D
IPR023858	RSAM_Hmdb	IPR034479	AhbC-like
IPR023862	CHP03960_rSAM	IPR034480	Heme_carboxy_lyase-like
IPR023863	rSAM_PTO1314	IPR034485	Anaerobic_Cys-type_sulfatase-m
IPR023867	Sulphatase_maturase_rSAM	IPR034491	Anaerob_Ser_sulfatase-maturase
IPR023868	7-CO-7-deazaGua_synth_put_Clo	IPR034497	Bacteriochlorophyll_C12_MT
IPR023874	DNA_rSAM_put	IPR034498	Bacteriochlorophyll_C8_MT
IPR023880	Benzylsucc_Synthase_activating	IPR034505	Coproporphyrinogen-III_oxidase
IPR023885	4Fe4S-binding_SPASM_dom	IPR034508	Spectinomycin_biosynthesis
IPR023886	QH-AmDH_gsu_maturation	IPR034514	ThnK-like
IPR023891	Pyrryls_PyIB	IPR034515	ThnL-like
IPR023897	Spore_PP_lyase	IPR034519	TunB-like
IPR023904	Pep_rSAM_mat_YydG	IPR034529	Fom3-like
IPR023912	YjjW_bact	IPR034530	HpnP-like
IPR023913	MftC	IPR034531	Methylation_of_yatakemycin
IPR023930	NirJ1	IPR034532	OxB-like
IPR023969	CHP04072_B12-bd/rSAM	IPR034534	Pyrimidine_methyltransferase
IPR023979	CHP04014_B12-bd/rSAM	IPR034547	Tte1186a_maturase
IPR023980	CHP04013_B12-bd/rSAM	IPR034556	tRNA_wybutosine-synthase

IPR023984	rSAM_ocin_1	IPR034557	ThrcA_tRNA_Methiotransferase
IPR023992	HemeD1_Synth_NirJ	IPR034559	Spore_PP_lyase_Clostridia
IPR023993	TYW1_archaea	IPR034560	Spore_PP_lyase_Bacilli
IPR023995	HemZ	IPR034687	ELP3-like
IPR024001	Cys-rich_pep_rSAM_mat_CcpM	IPR038135	Methylthiotransferase_N_sf
IPR024007	FeFe-hyd_mat_HydG	IPR039661	ELP3
IPR024016	CHP04064_rSAM	IPR040072	Methyltransferase_A
IPR024017	Pep_cycl_rSAM	IPR040074	BssD/PflA/YjjW
IPR024018	CHP04083_rSAM	IPR040081	CndI-like
IPR024021	FeFe-hyd_HydE_rSAM	IPR040082	GenK-like
IPR024023	rSAM_paired_HxsB	IPR040085	MJ0674-like
IPR024025	SCIFF_rSAM_maturase	IPR040086	MJ0683-like
IPR024032	rSAM_paired_HxsC	IPR040087	MJ0021-like
IPR024177	Biotin_synthase	IPR040088	MJ0103-like
IPR024521	DUF3641	IPR041582	RimO_TRAM
IPR024560	UPF0313_C	IPR045375	Put_radical_SAM-like_N
IPR024924	7-CO-7-deazaguanine_synth-like	IPR045567	CofH/MnqC-like_C
IPR025895	LAM_C_dom	IPR045784	Radical_SAM_N2
IPR026322	Geopep_mat_rSAM	PF04055	Radical_SAM
IPR026332	HutW	PF06969	HemN_C
IPR026335	SAM_SPASM_FxsB	PF08497	Radical_SAM_N
IPR026344	SCM_rSAM_ScmE	PF12345	DUF3641
IPR026346	SCM_rSAM_ScmF	PF13186	SPASM
IPR026351	rSAM_SeCys	PF16199	Radical_SAM_C
IPR026357	rSAM/SPASM_prot_GRRM_system	PF16881	LIAS_N
IPR026401	CXXX_matur	PF19238	Radical_SAM_2
IPR026404	rSAM_w_lipo	PF19288	CofH_C
IPR026407	SAM_GG-Bacter	PF19864	Radical_SAM_N2
IPR026412	rSAM_Cxxx_rpt		

References

- [1] Zallot R, Oberg N, Gerlt JA. The EFI Web Resource for Genomic Enzymology Tools: Leveraging Protein, Genome, and Metagenome Databases to Discover Novel Enzymes and Metabolic Pathways. *Biochemistry*. 2019;58:4169-82.
- [2] Oberg N, Precord TW, Mitchell DA, Gerlt JA. RadicalSAM.org: A Resource to Interpret Sequence-Function Space and Discover New Radical SAM Enzyme Chemistry. *ACS Bio Med Chem Au*. 2022;2:22-35.
- [3] Holliday GL, Akiva E, Meng EC, Brown SD, Calhoun S, Pieper U, et al. Atlas of the Radical SAM Superfamily: Divergent Evolution of Function Using a "Plug and Play" Domain. *Methods Enzymol*. 2018;606:1-71.