SUPPLEMENTAL TUTORIALS

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

Nils Oberg[1], Rémi Zallot[2,3], and John A. Gerlt[1,4,5*]

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

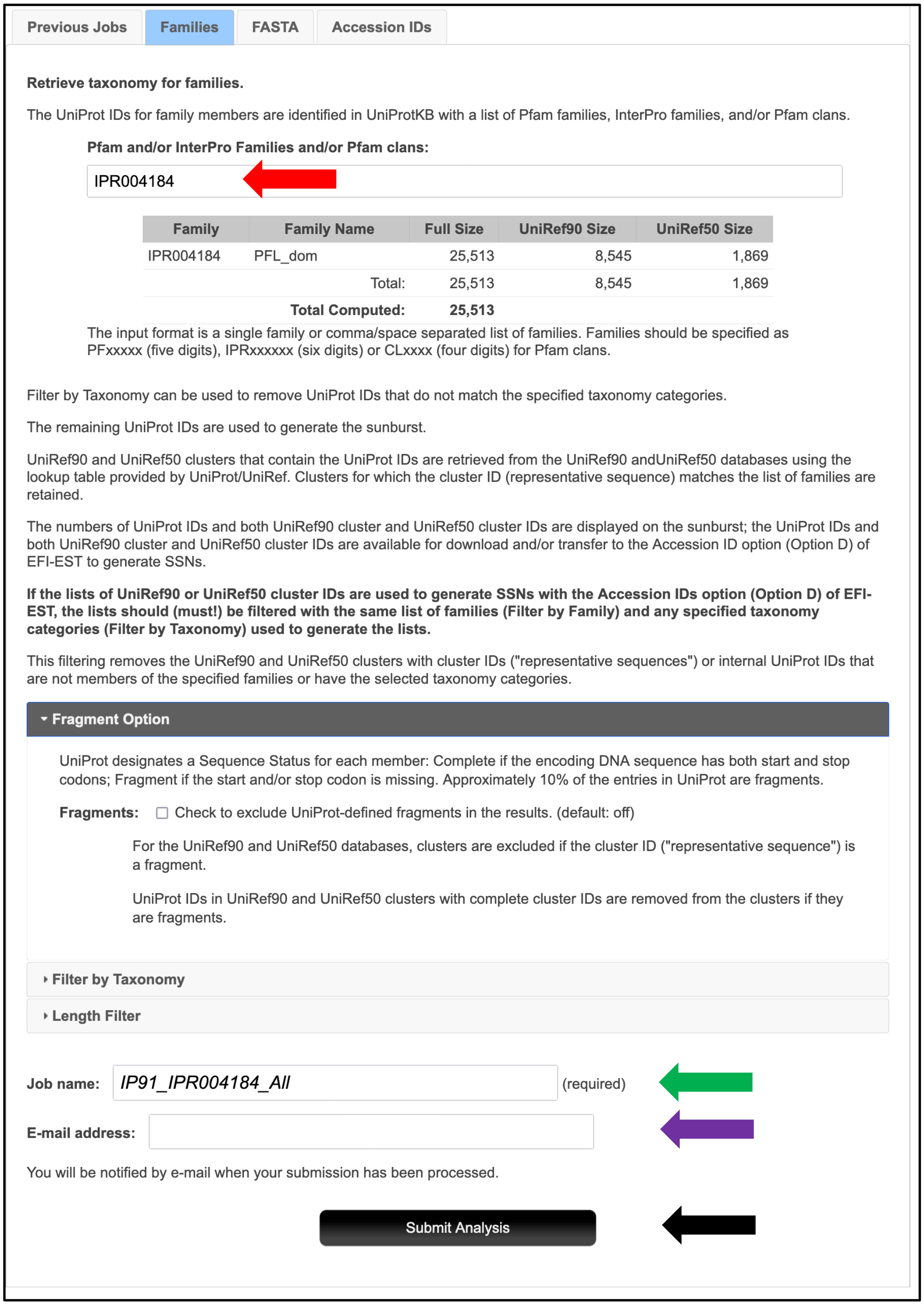
[2]Department of Chemistry, [3]Manchester Institute of Biotechnology, The University of Manchester, 131 Princess Street, Manchester M1 7DN, UK

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).



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 andUniRef50 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.

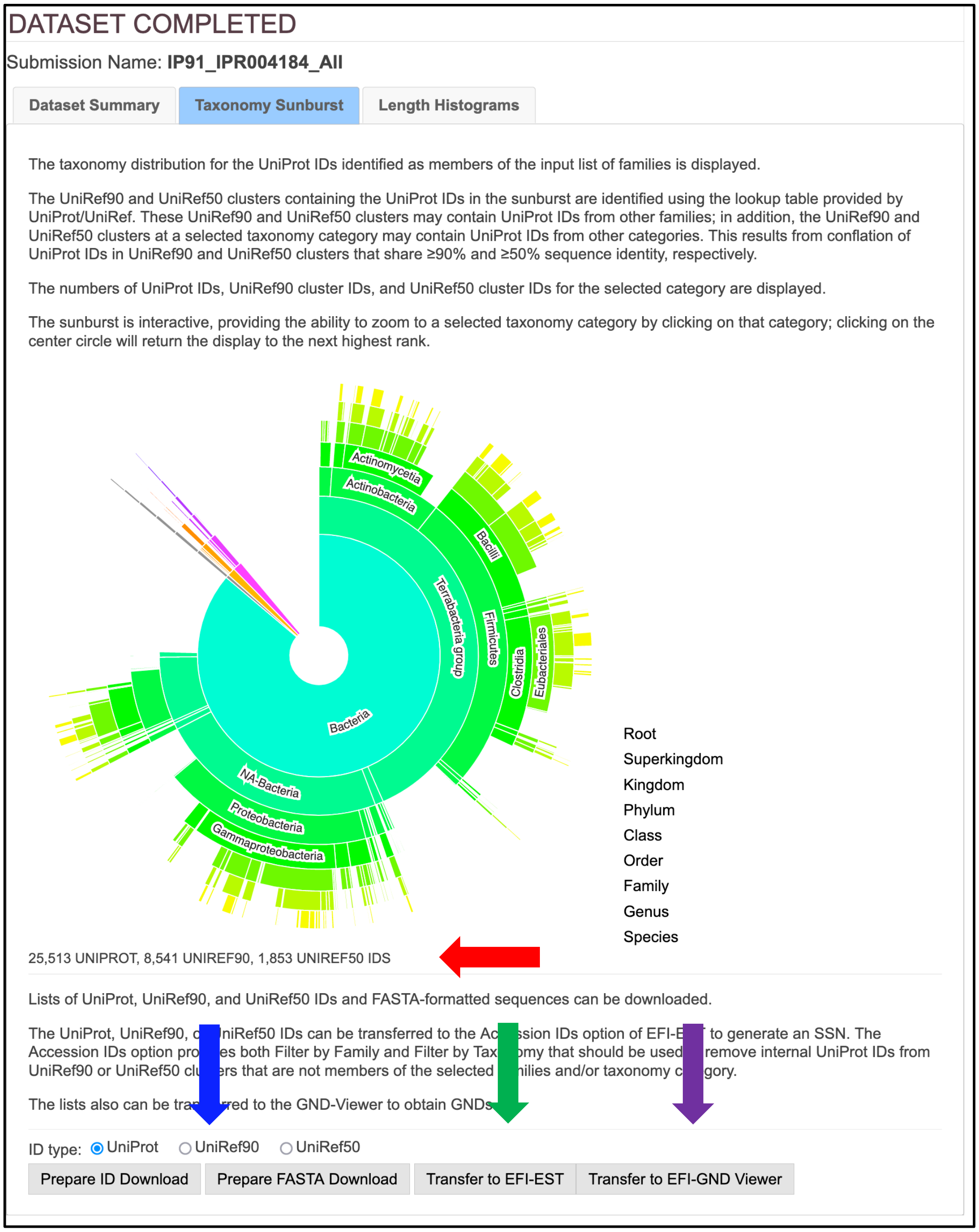
UniProt IDs in UniRef90 and UniRef50 clusters with complete cluster IDs are removed from the clusters if they are fragments.

Job name: *IP91_IPR004184_All* (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_IPR004184_All**

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 ≥90% and ≥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.

25,513 UNIPROT, 8,541 UNIREF90, 1,853 UNIREF50 IDS

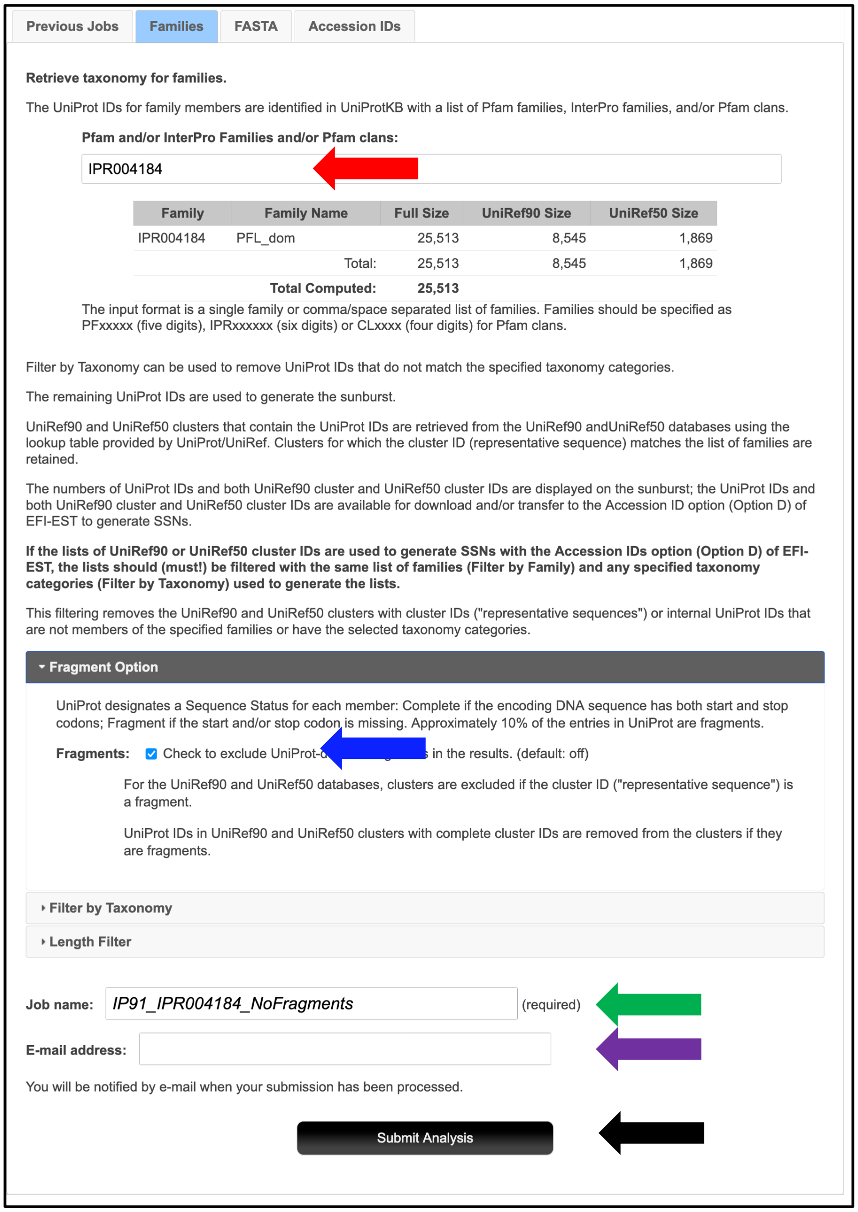
Lists of UniProt, UniRef90, and UniRef50 IDs and FASTA-formatted sequences can be downloaded.

The UniProt, UniRef90, o niRef50 IDs can be transferred to the Ac ssion IDs option of EFI-E to generate an SSN. The Accession IDs option pro es both Filter by Family and Filter by Tax my that should be used remove internal UniProt IDs from UniRef90 or UniRef50 clu ers that are not members of the selected ilies and/or taxonomy c gory.

The lists also can be tra red to the GND-Viewer to obtain GNDs

ID type: ● UniProt ○ UniRef90 ○ UniRef50

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).



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 andUniRef50 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

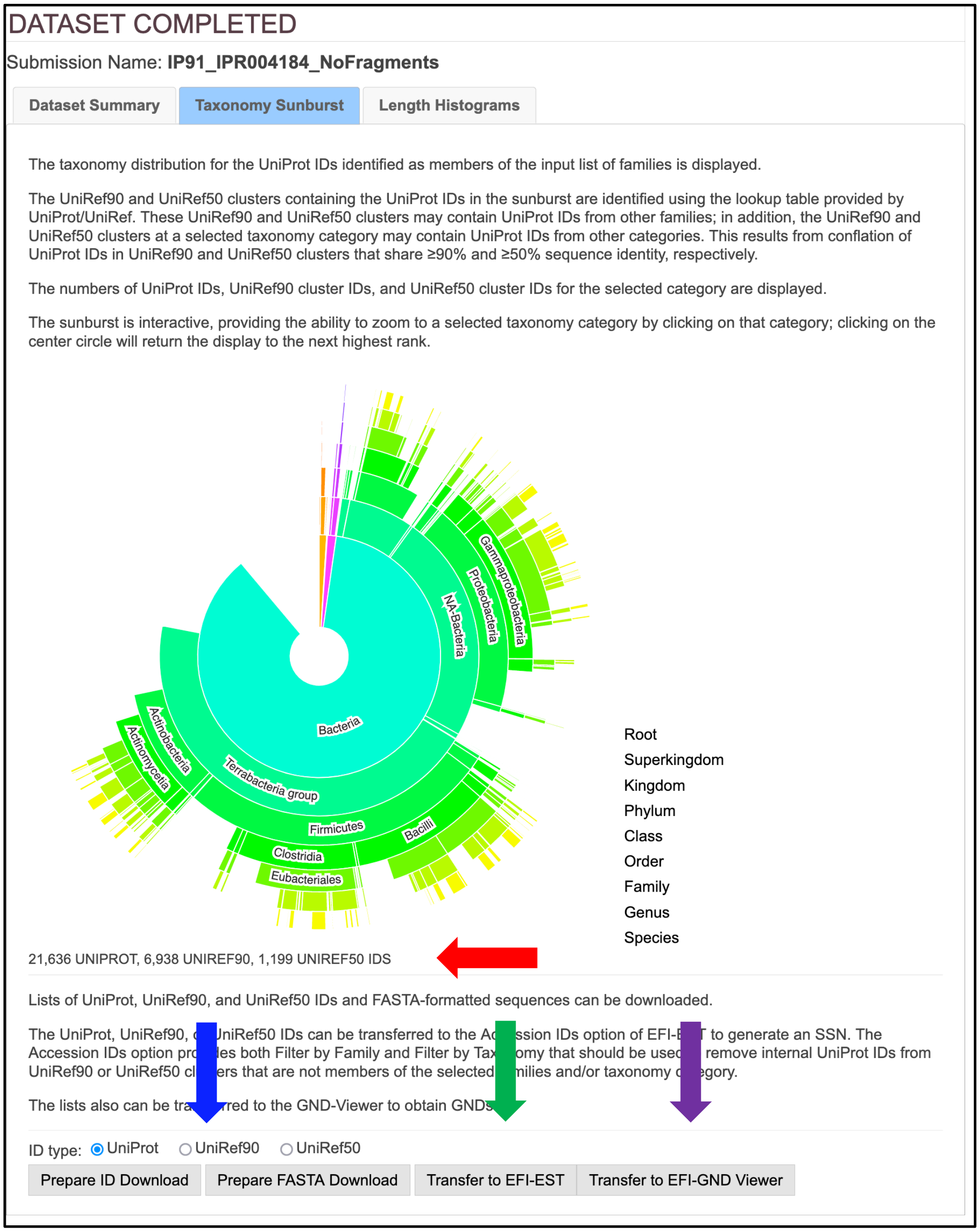
Length Filter

Job name: *IP91_IPR004184_NoFragments* (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_IPR004184_NoFragments**

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 ≥90% and ≥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.

21,636 UNIPROT, 6,938 UNIREF90, 1,199 UNIREF50 IDS

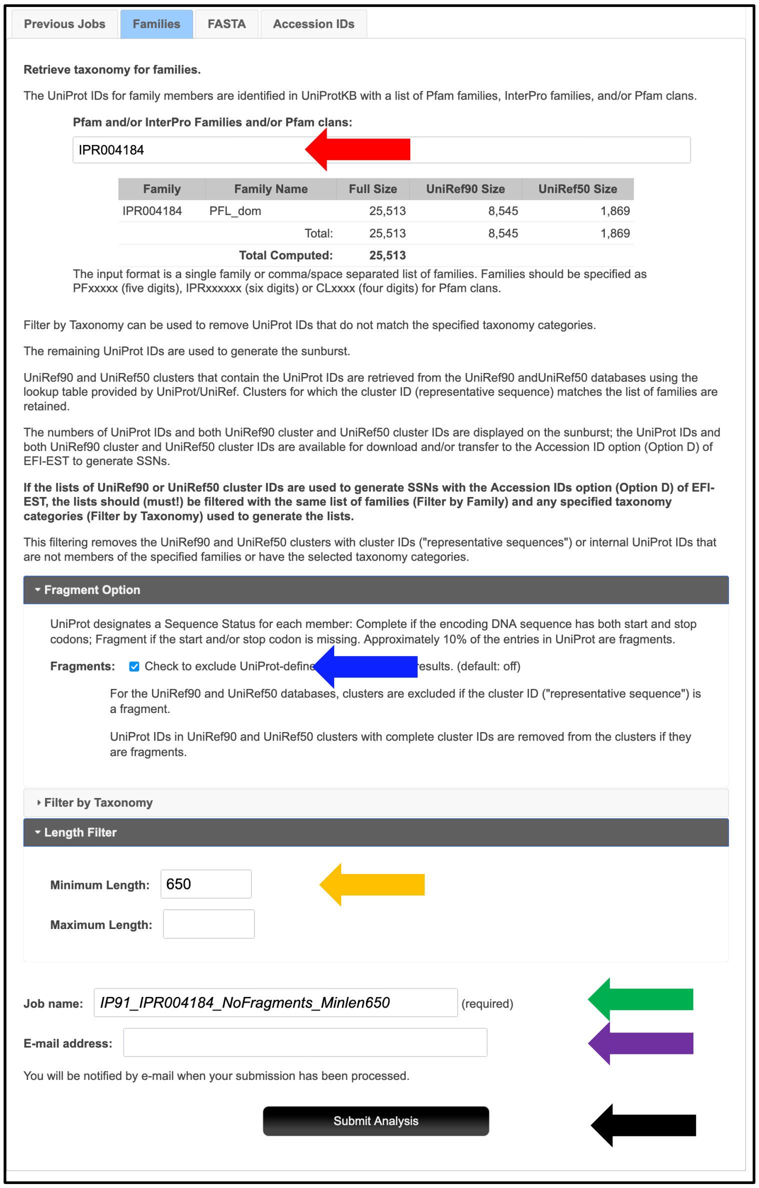
Lists of UniProt, UniRef90, and UniRef50 IDs and FASTA-formatted sequences can be downloaded.

The UniProt, UniRef90, or 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 GNDs.

ID type: ● UniProt ○ UniRef90 ○ UniRef50

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).



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 andUniRef50 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-define esults. (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: 650

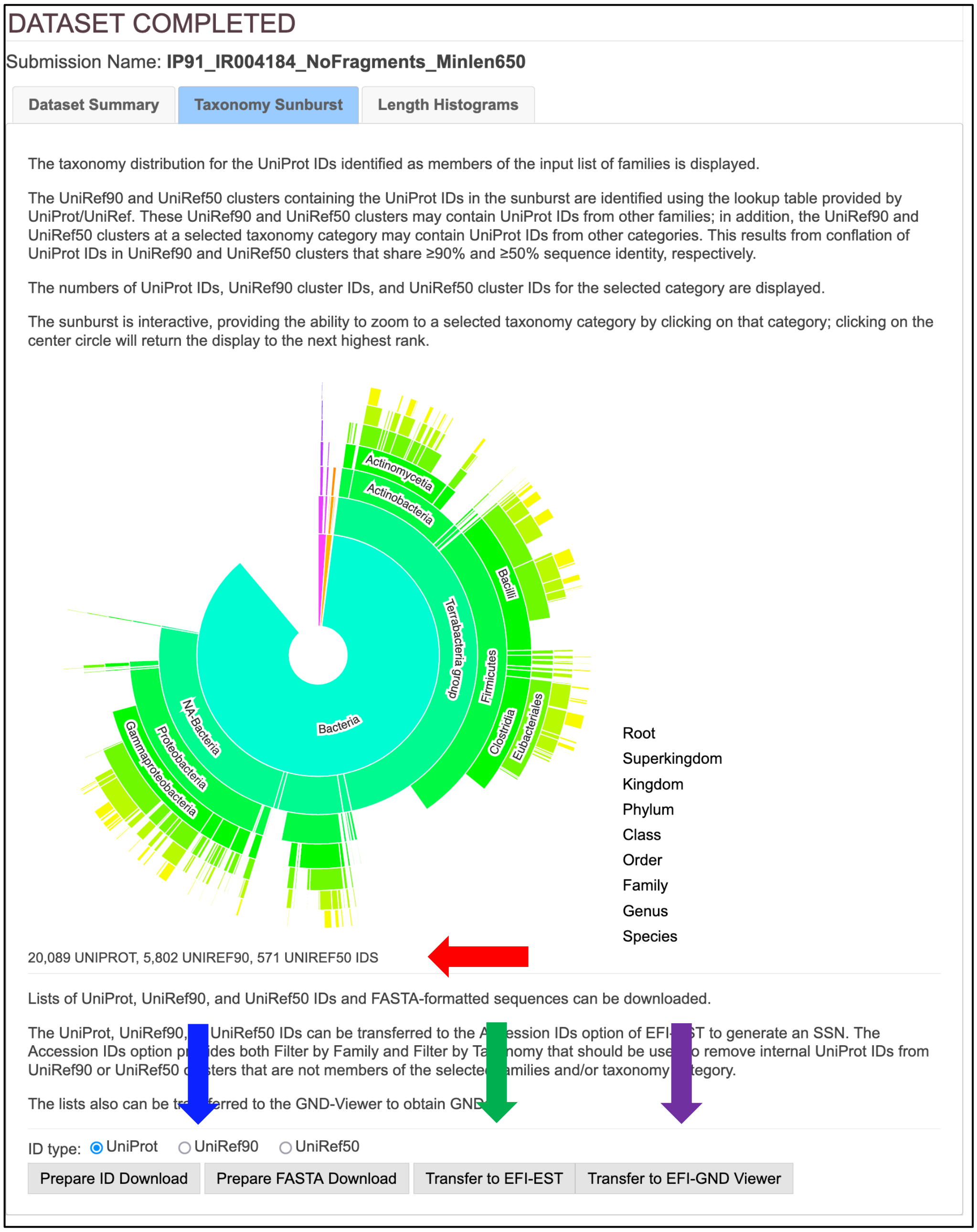
Maximum Length:

Job name: *IP91_IPR004184_NoFragments_Minlen650* (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**

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 ≥90% and ≥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.

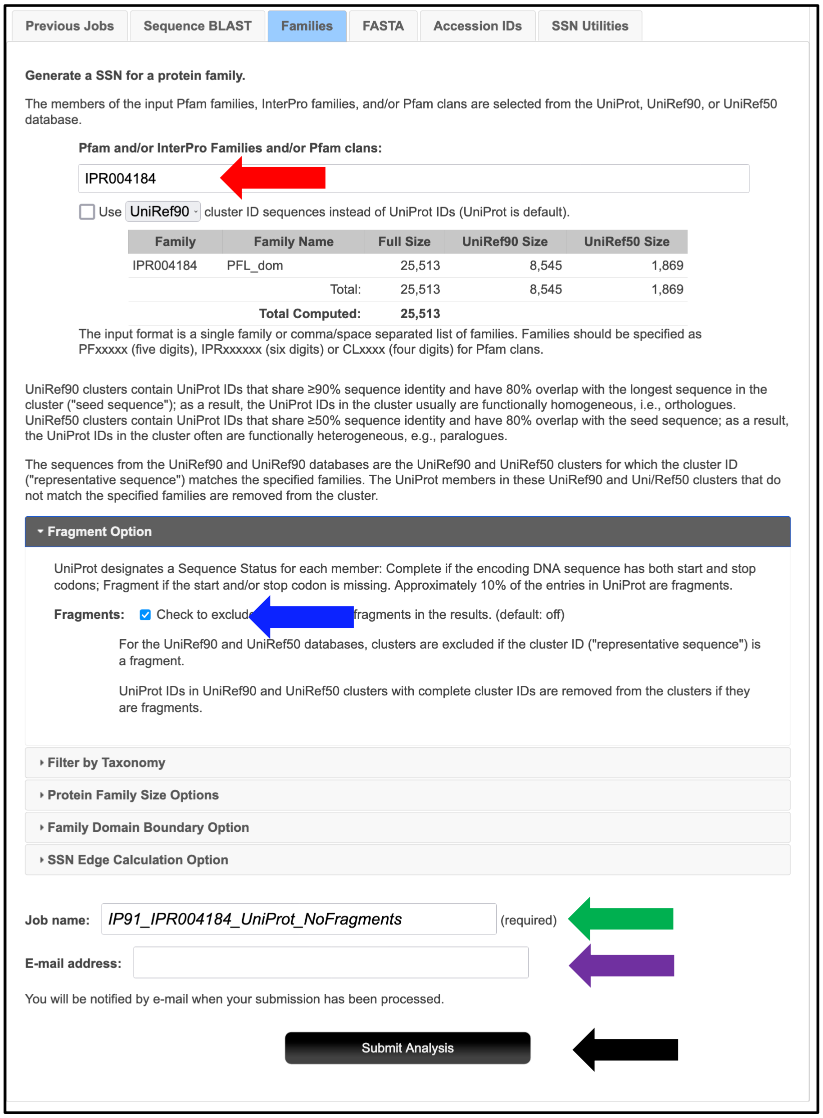
The UniProt, UniRef90, and UniRef50 IDs can be transferred to the Accession IDs option of EFI-EST to generate an SSN. The Accession IDs option provides both Filter by Family and Filter by Taxonomy that should be used to remove internal UniProt IDs from UniRef90 or UniRef50 clusters that are not members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

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

ID type: ● UniProt ○ UniRef90 ○ UniRef50

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).



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 ≥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 UniRef90 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 Uni/Ref50 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 exclud 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.

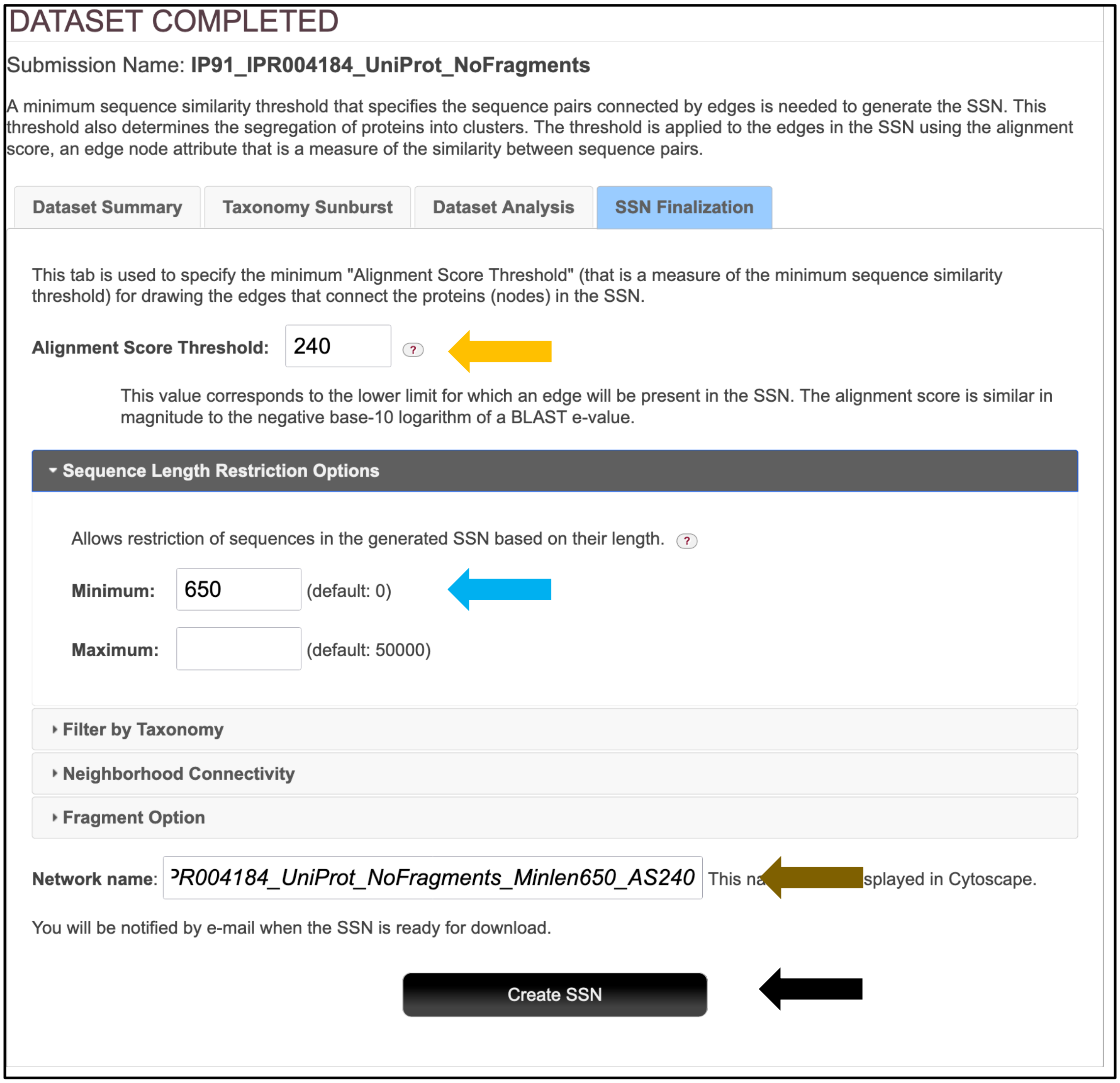
Filter by Taxonomy
Protein Family Size Options
Family Domain Boundary Option
SSN Edge Calculation Option

Job name: *IP91_IPR004184_UniProt_NoFragments* (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.

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: 240

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.

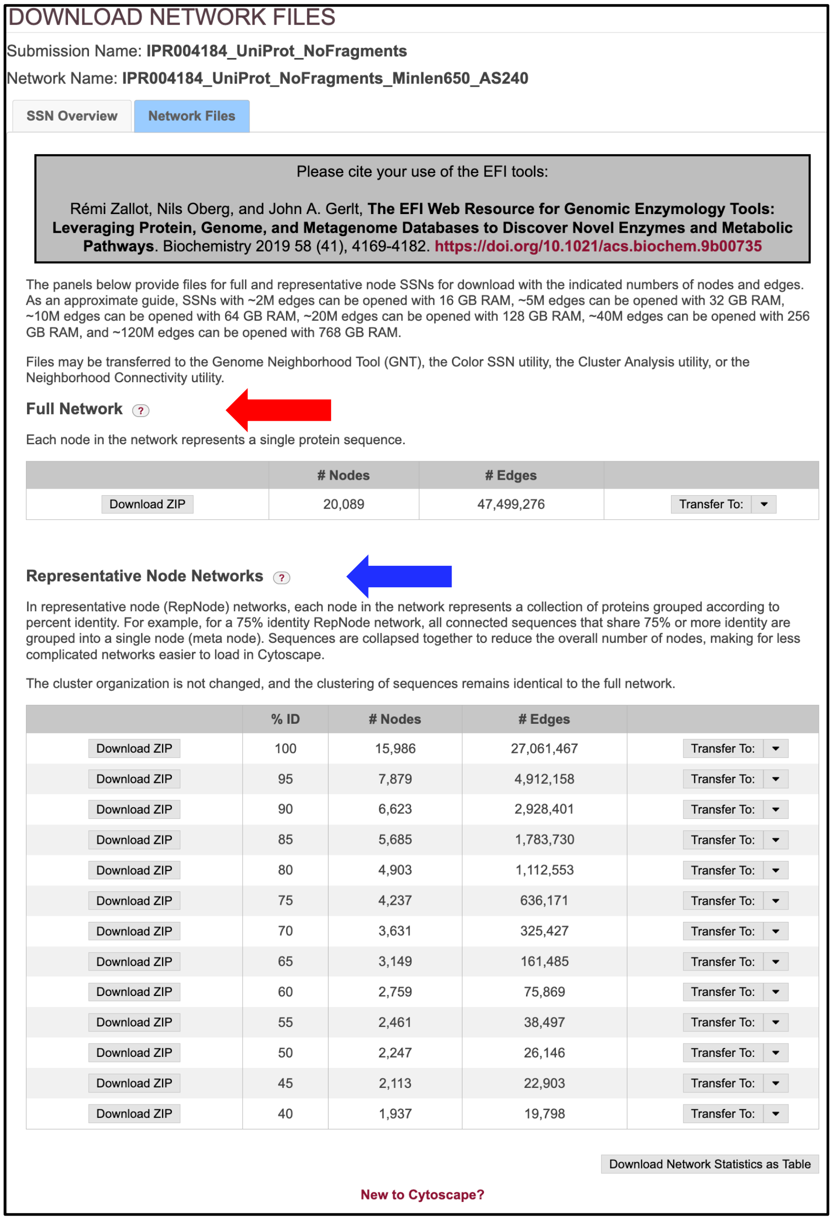
Minimum: 650 (default: 0)

Maximum: (default: 50000)

Network name: ᴾR004184_UniProt_NoFragments_Minlen650_AS240 This na splayed 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**

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.

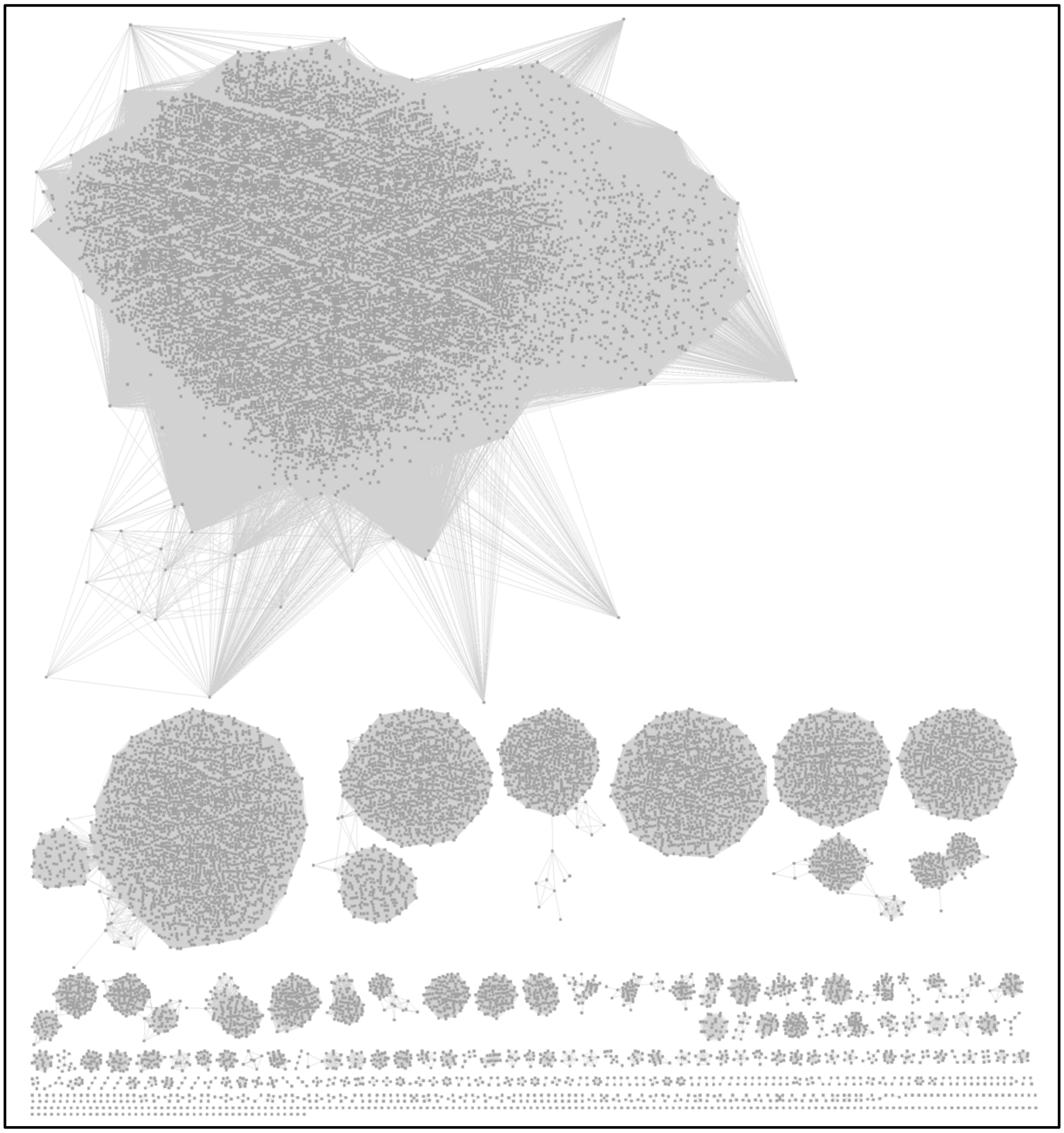
	# Nodes	# Edges	
Download ZIP	20,089	47,499,276	Transfer To: ▾

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.

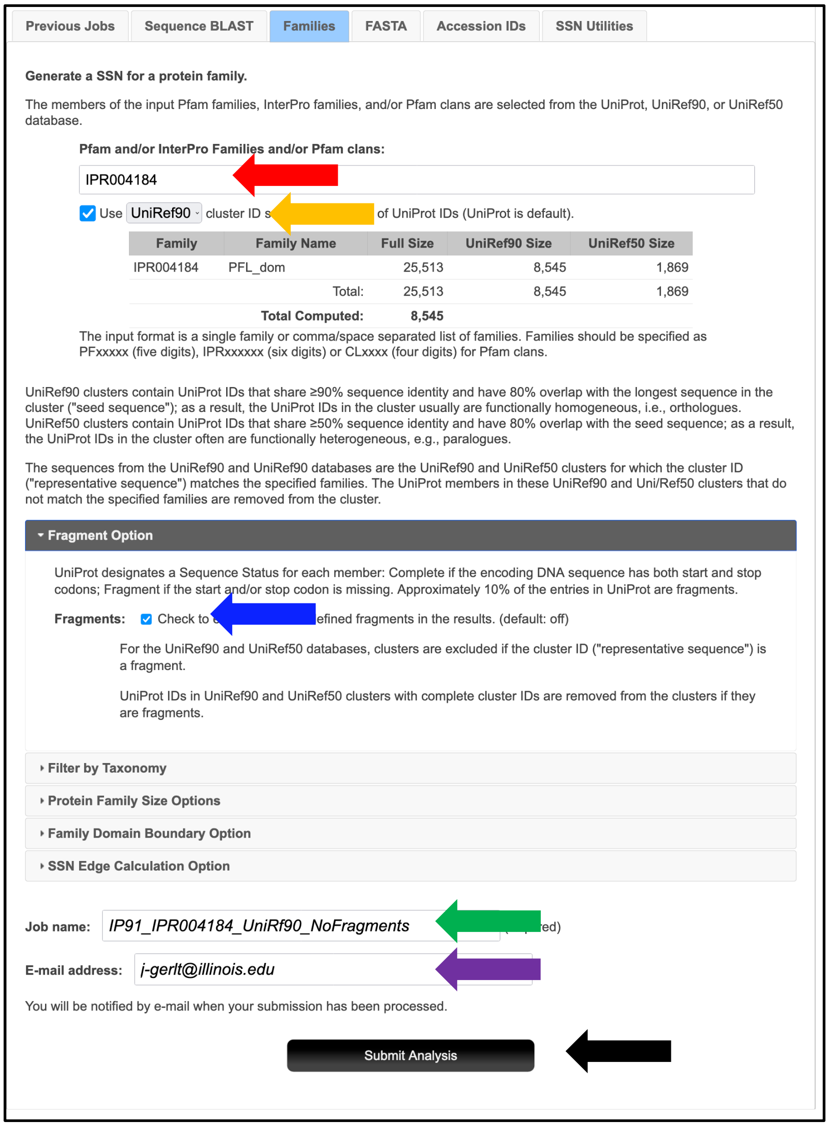
	% ID	# Nodes	# Edges	
Download ZIP	100	15,986	27,061,467	Transfer To: ▾
Download ZIP	95	7,879	4,912,158	Transfer To: ▾
Download ZIP	90	6,623	2,928,401	Transfer To: ▾
Download ZIP	85	5,685	1,783,730	Transfer To: ▾
Download ZIP	80	4,903	1,112,553	Transfer To: ▾
Download ZIP	75	4,237	636,171	Transfer To: ▾
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: ▾



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).



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 s 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 ≥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 UniRef90 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 Uni/Ref50 clusters that do not match the specified families are removed from the cluster.

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 efined 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.

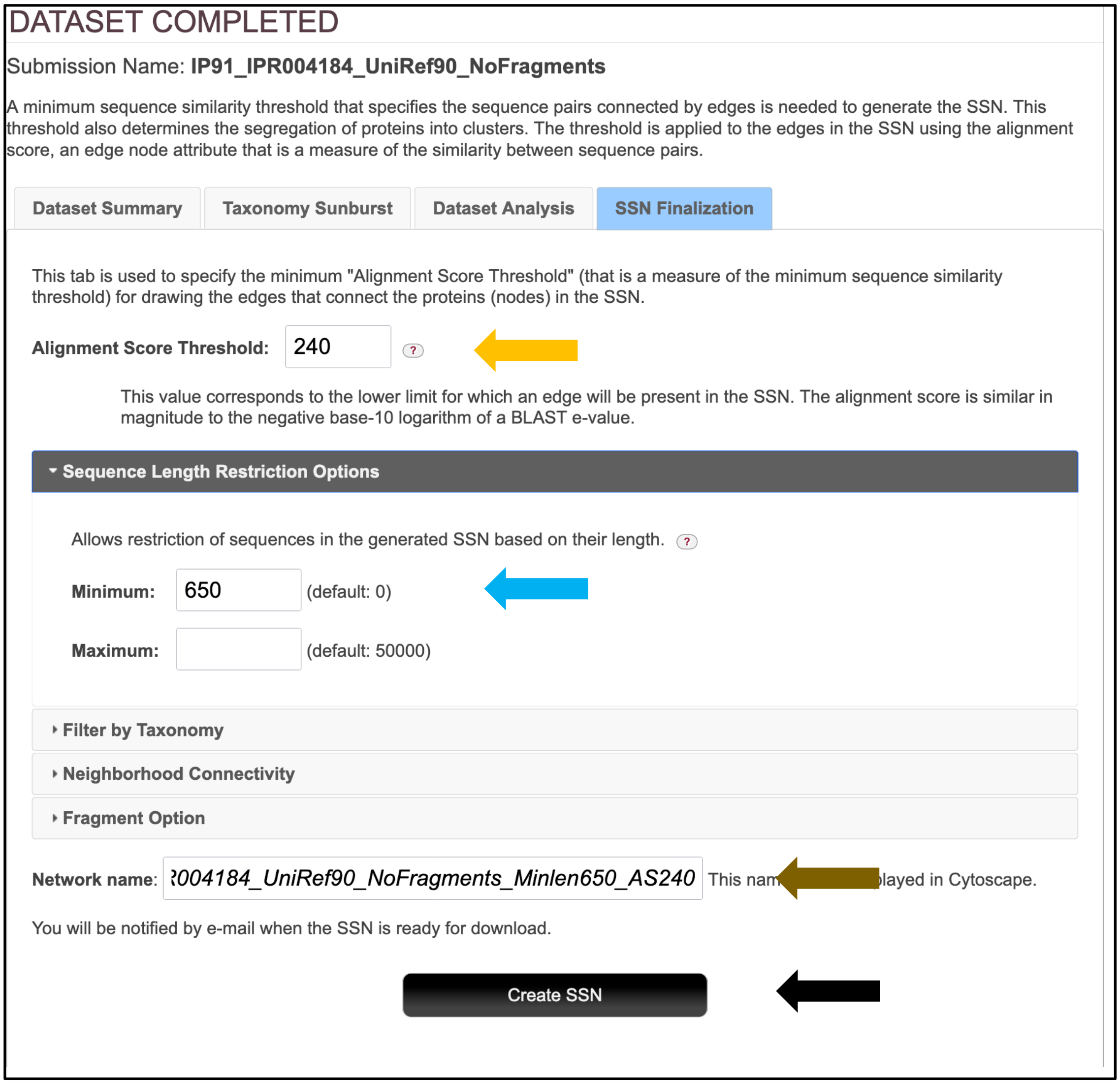
▸ Filter by Taxonomy
▸ Protein Family Size Options
▸ Family Domain Boundary Option
▸ SSN Edge Calculation Option

Job name: *IP91_IPR004184_UniRf90_NoFragments* ed)

E-mail address: *j-gerlt@illinois.edu*

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.

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: 240

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.

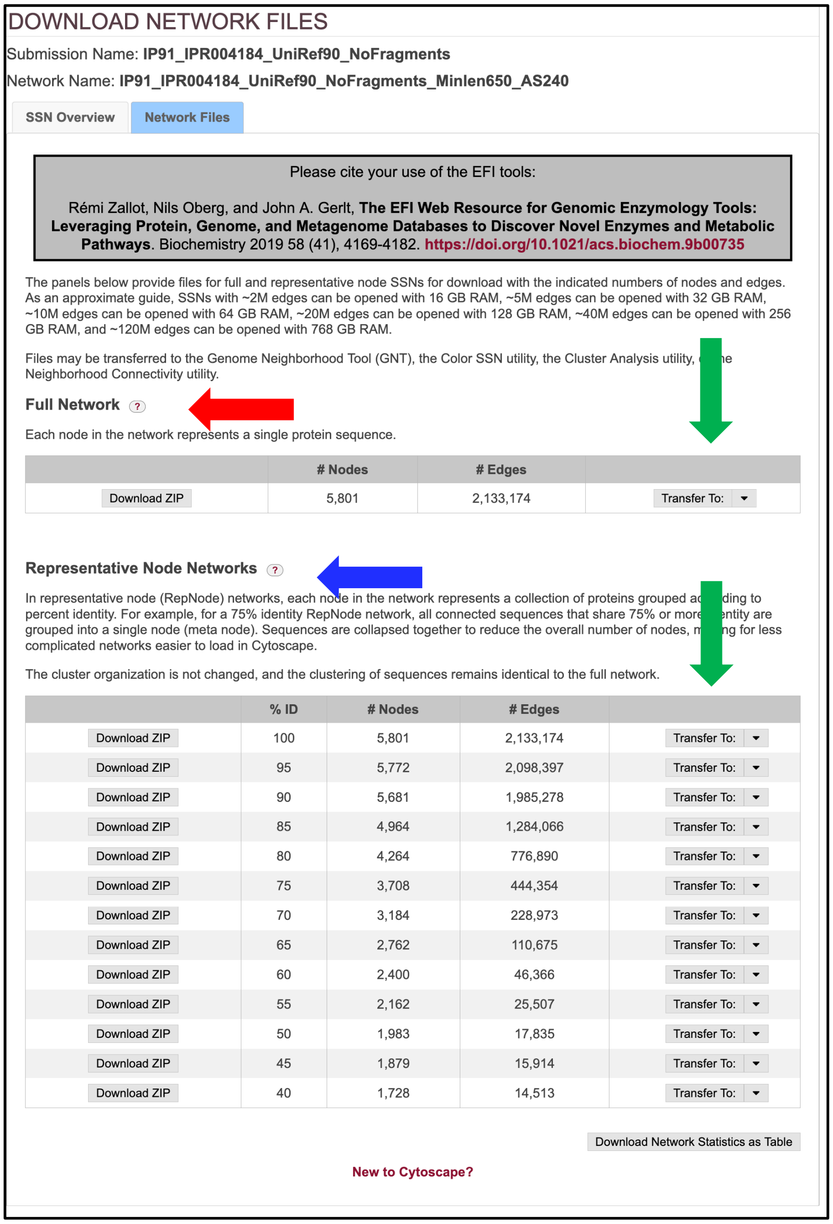
Minimum: 650 (default: 0)

Maximum: (default: 50000)

Network name: ?004184_UniRef90_NoFragments_Minlen650_AS240 This nam layed 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). 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**

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	5,801	2,133,174	Transfer To: ▾

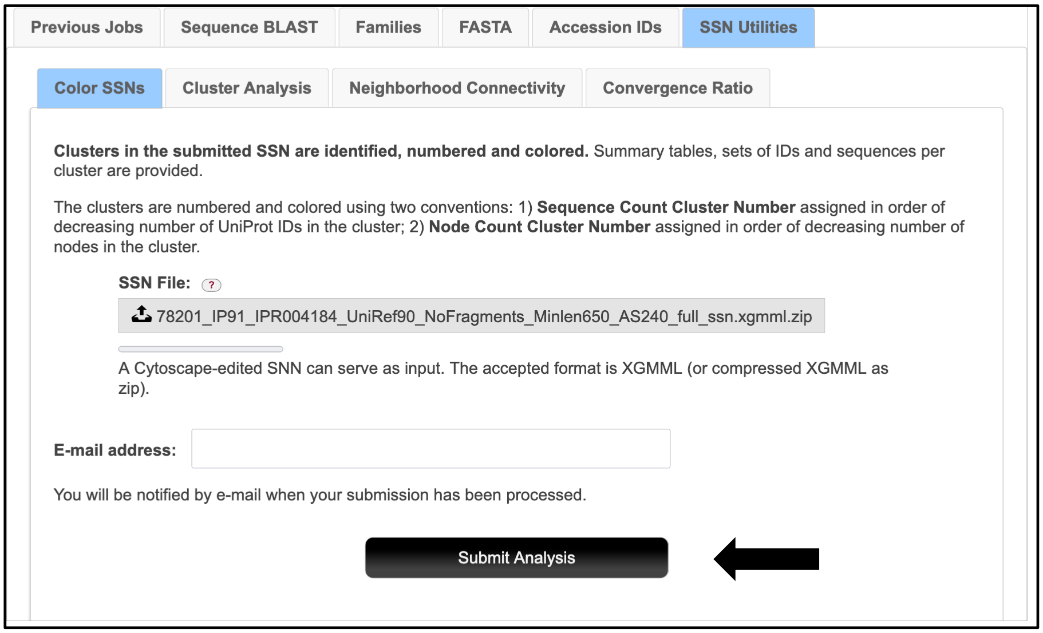
Representative Node Networks ?

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: ▾

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).



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.

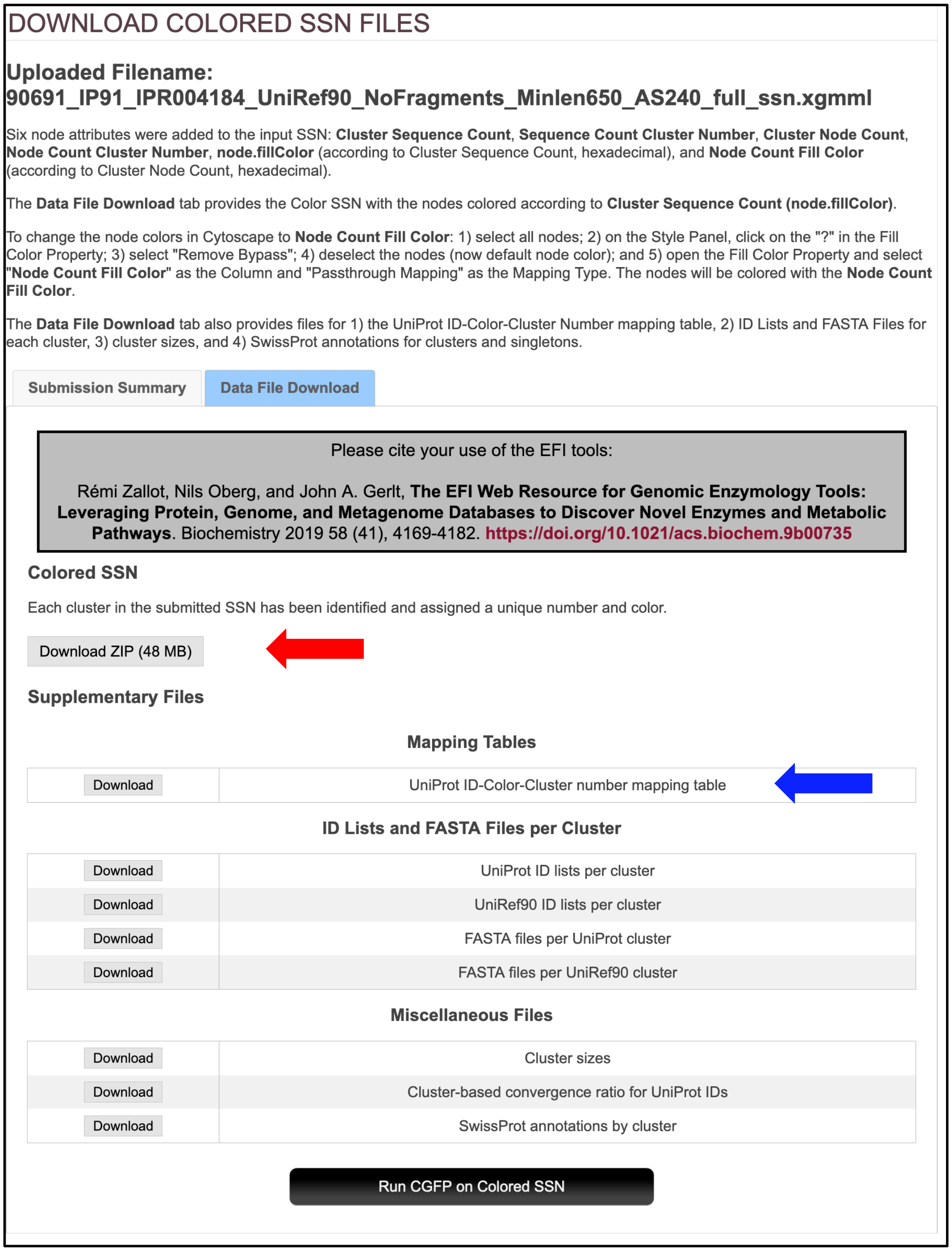
SSN File: 78201_IP91_IPR004184_UniRef90_NoFragments_Minlen650_AS240_full_ssn.xgmml.zip

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.

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.

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.

Supplementary Files

Mapping Tables

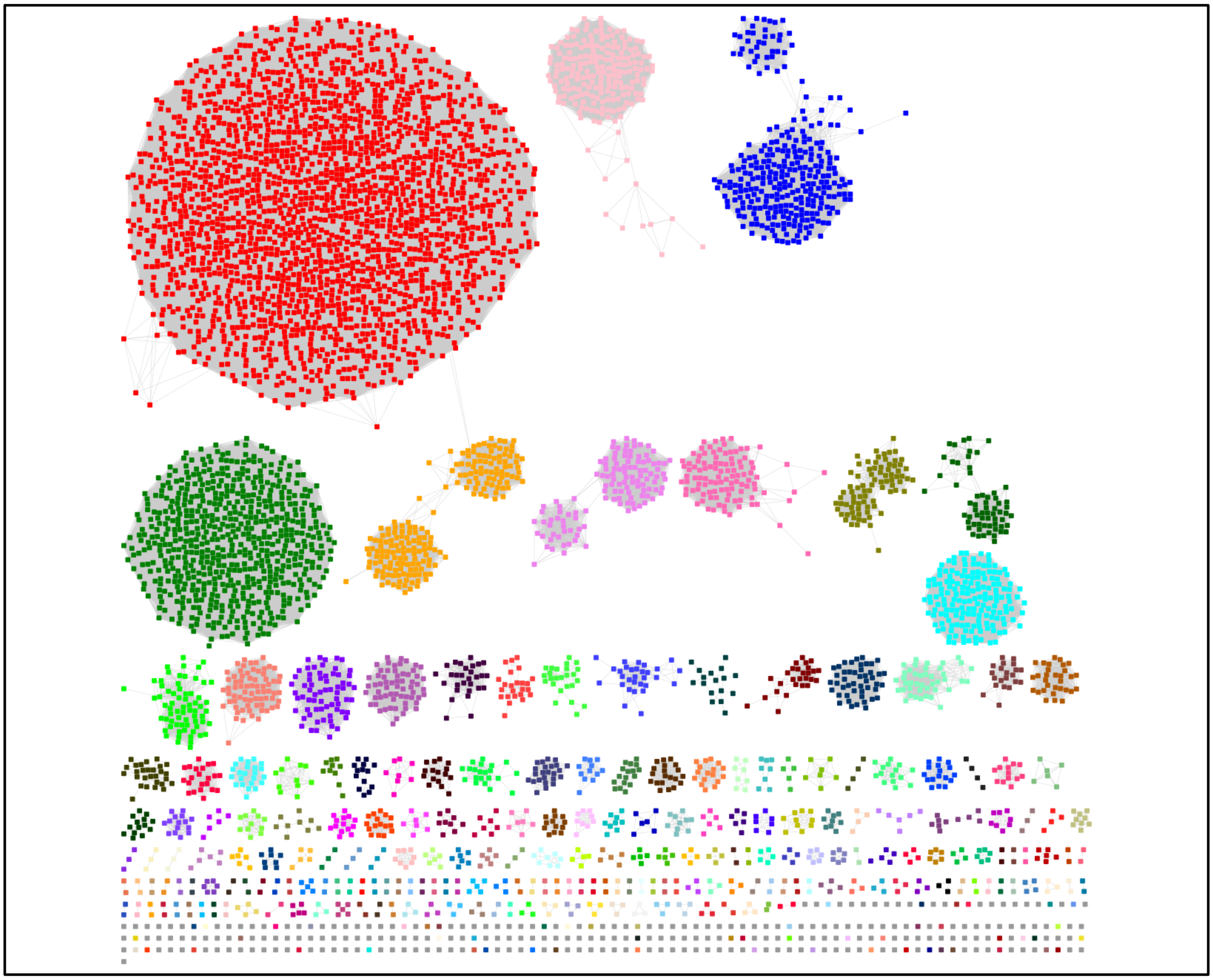
Download	UniProt ID-Color-Cluster number mapping table

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

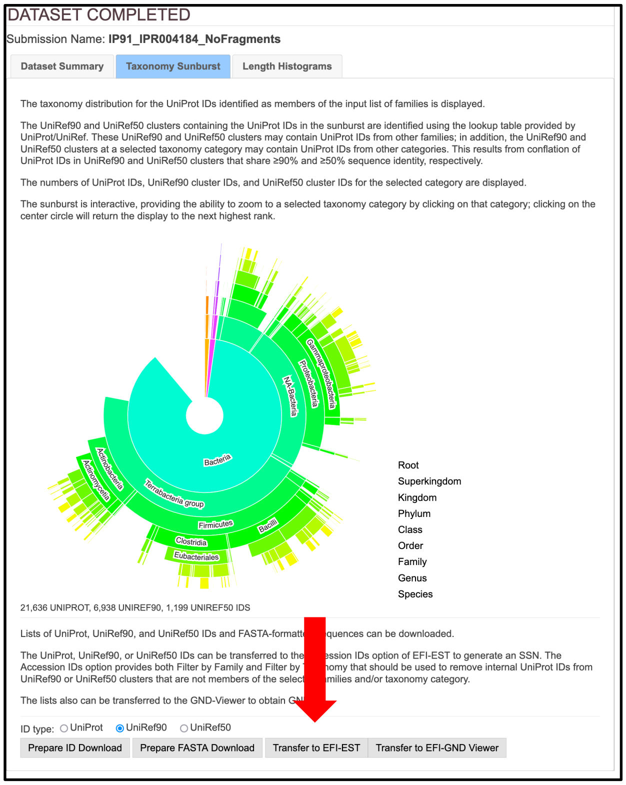


UniRef90 Cluster SSN for the GRE Superfamily. As described in the previous sections, the UniRef90 cluster SSN for the GRE superfamily was generated using an alignment score threshold of 240 and a minimum length of 650 residues. The nodes were colored using the **Color SSNs** utility. The SSN contains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



DATASET COMPLETED

Submission Name: **IP91_IPR004184_NoFragments**

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 ≥90% and ≥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.

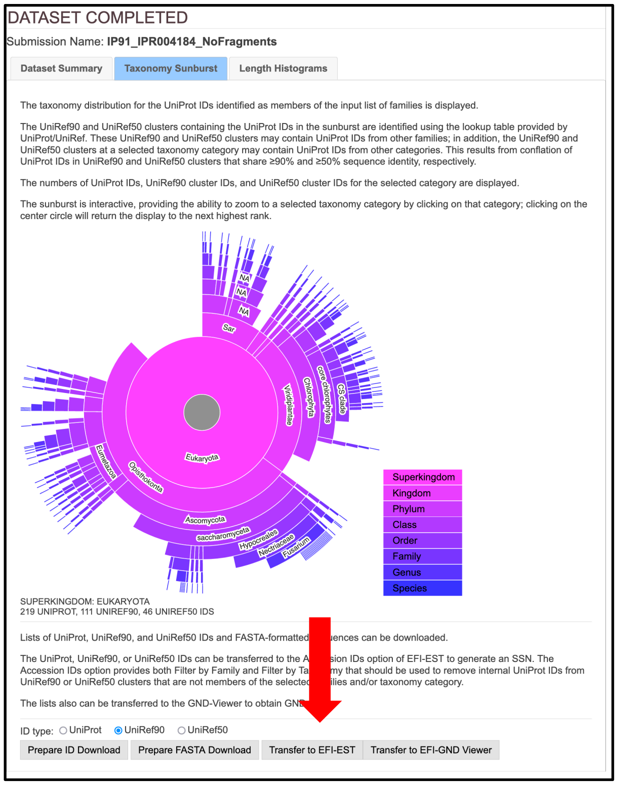
21,636 UNIPROT, 6,938 UNIREF90, 1,199 UNIREF50 IDS

Lists of UniProt, UniRef90, and UniRef50 IDs and FASTA-formatte quences can be downloaded.

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

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

ID type: ○ UniProt ● UniRef90 ○ UniRef50



DATASET COMPLETED

Submission Name: **IP91_IPR004184_NoFragments**

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 ≥90% and ≥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.

SUPERKINGDOM: EUKARYOTA
219 UNIPROT, 111 UNIREF90, 46 UNIREF50 IDS

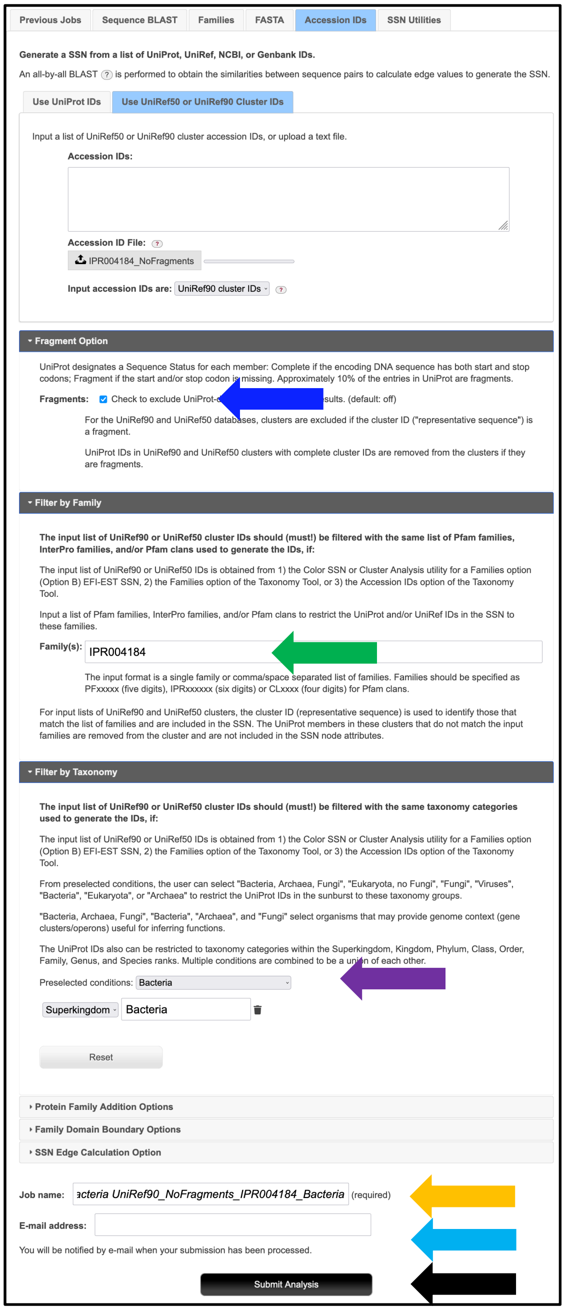
Lists of UniProt, UniRef90, and UniRef50 IDs and FASTA-formatted uences can be downloaded.

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

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

ID type: ○ UniProt ● UniRef90 ○ UniRef50

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).



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.

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

Accession IDs:

Accession ID File: ?

IPR004184_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-d sults. (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) EFI-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): IPR004184

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) EFI-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: Bacteria

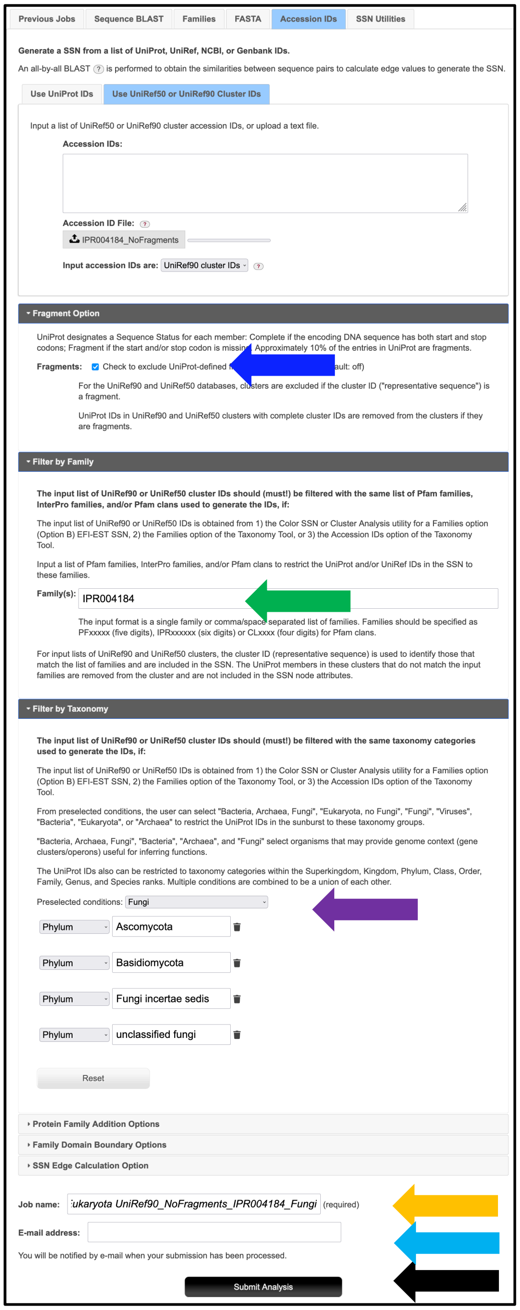
Superkingdom Bacteria

Reset
Protein Family Addition Options
Family Domain Boundary Options
SSN Edge Calculation Option

Job name: acteria UniRef90_NoFragments_IPR004184_Bacteria (required)

E-mail address:

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



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.

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

Accession IDs:

Accession ID File: ?

IPR004184_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 f ault: 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) EFI-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): IPR004184

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) EFI-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

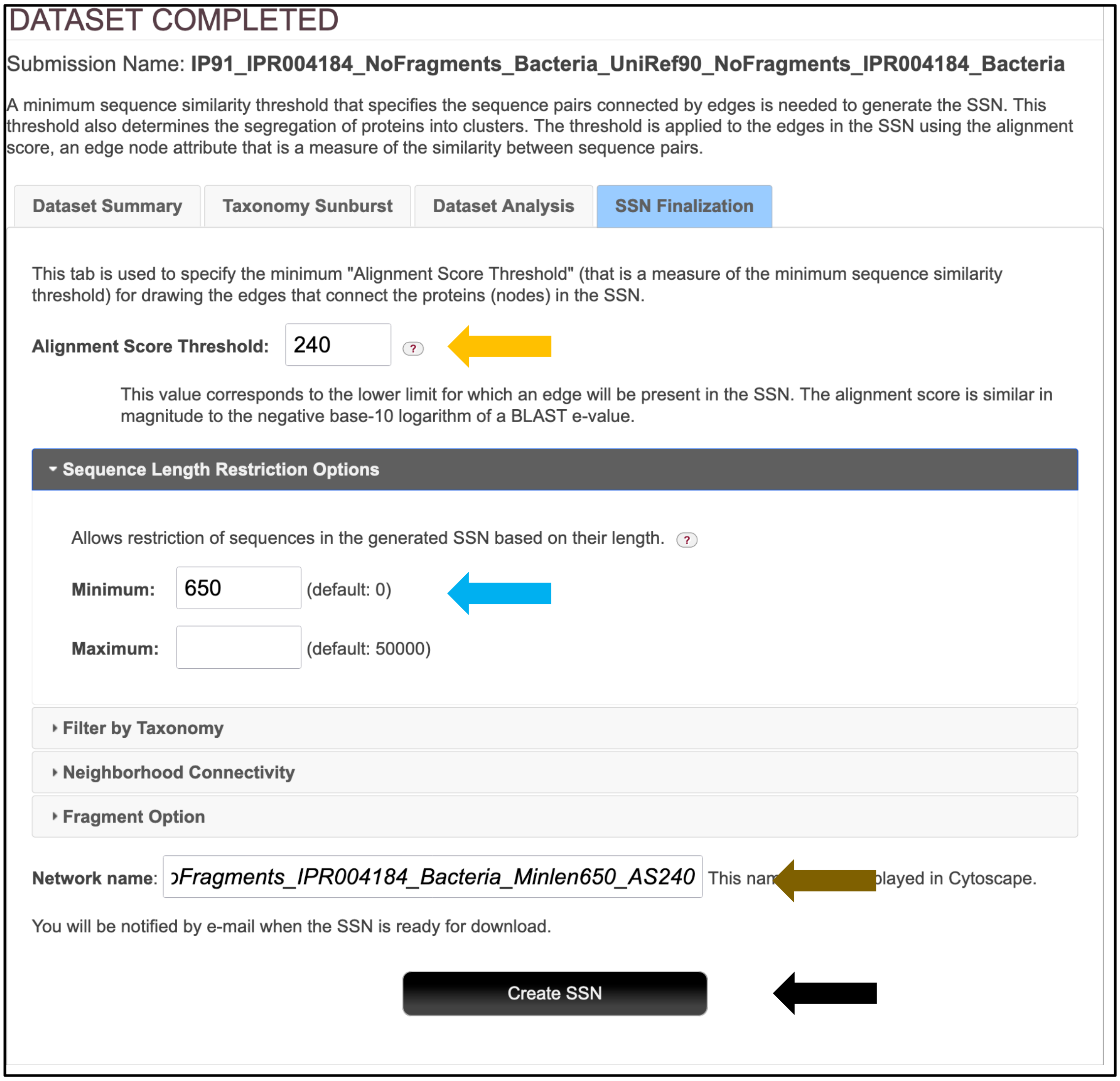
Reset
Protein Family Addition Options
Family Domain Boundary Options
SSN Edge Calculation Option

Job name: :ukaryota UniRef90_NoFragments_IPR004184_Fungi (required)

E-mail address:

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

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: 240

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.

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

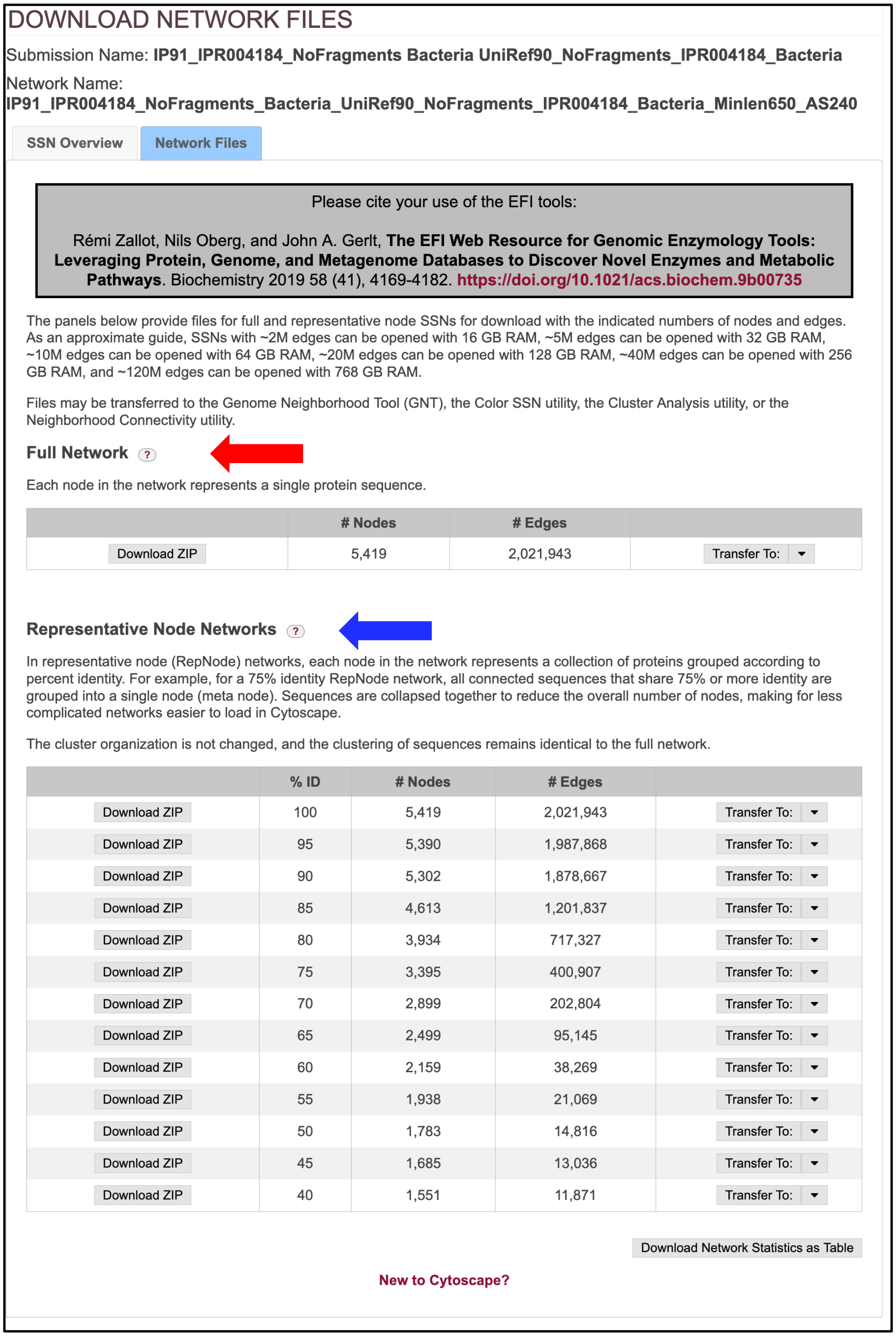
Minimum: 650 (default: 0)

Maximum: (default: 50000)

Network name: *ɔFragments_IPR004184_Bacteria_Minlen650_AS240* This nam 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

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.

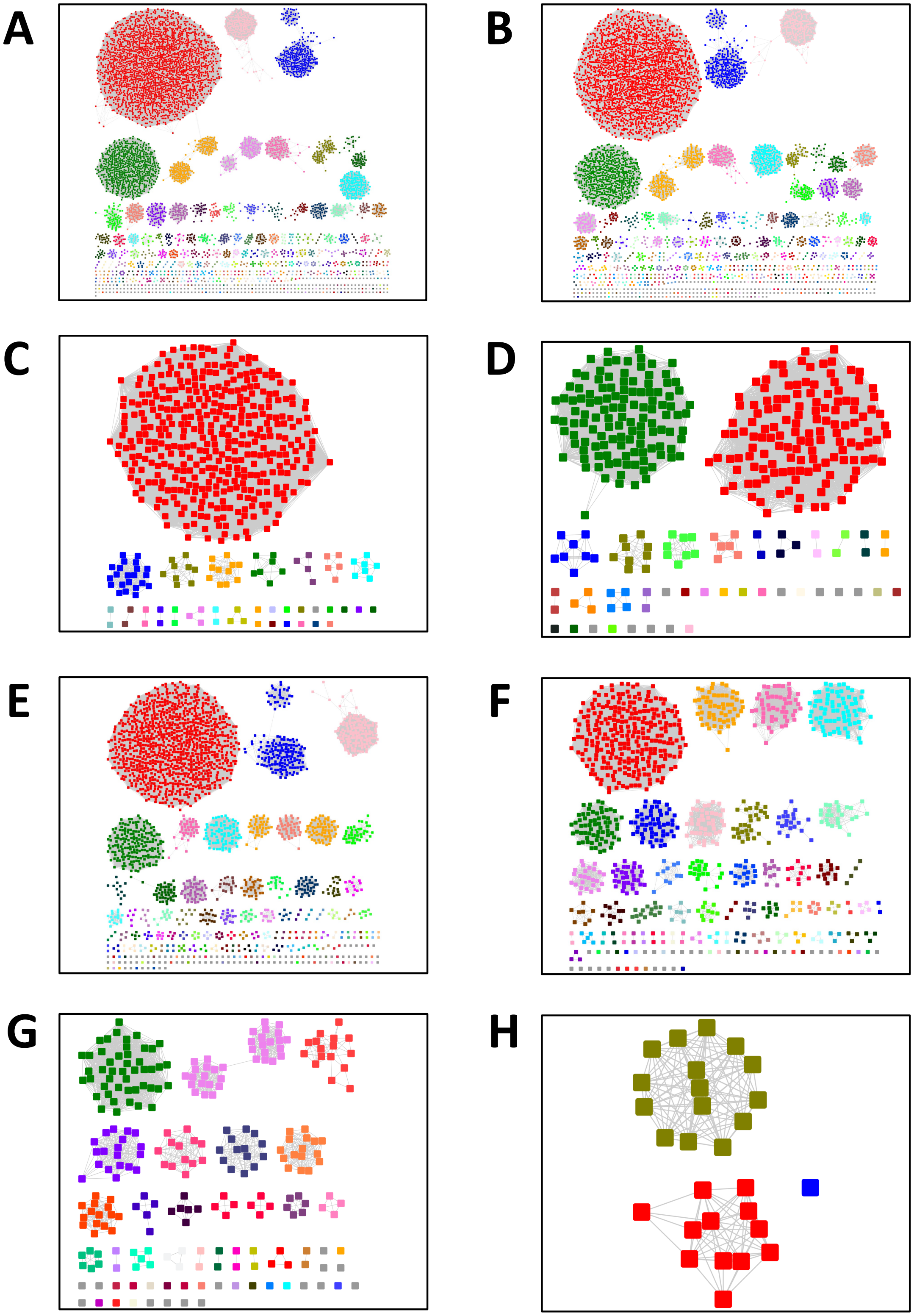
	# Nodes	# Edges	
Download ZIP	5,419	2,021,943	Transfer To: ▾

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.

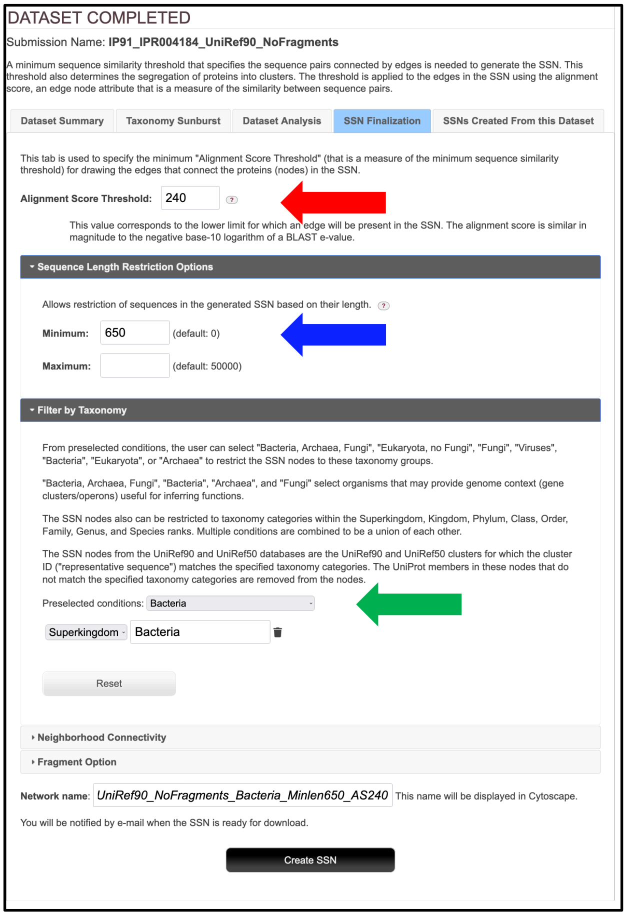
	% ID	# Nodes	# Edges	
Download ZIP	100	5,419	2,021,943	Transfer To: ▾
Download ZIP	95	5,390	1,987,868	Transfer To: ▾
Download ZIP	90	5,302	1,878,667	Transfer To: ▾
Download ZIP	85	4,613	1,201,837	Transfer To: ▾
Download ZIP	80	3,934	717,327	Transfer To: ▾
Download ZIP	75	3,395	400,907	Transfer To: ▾
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: ▾


A
B
C
D
E
F
G
H

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.

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: 240

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: 650 (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.

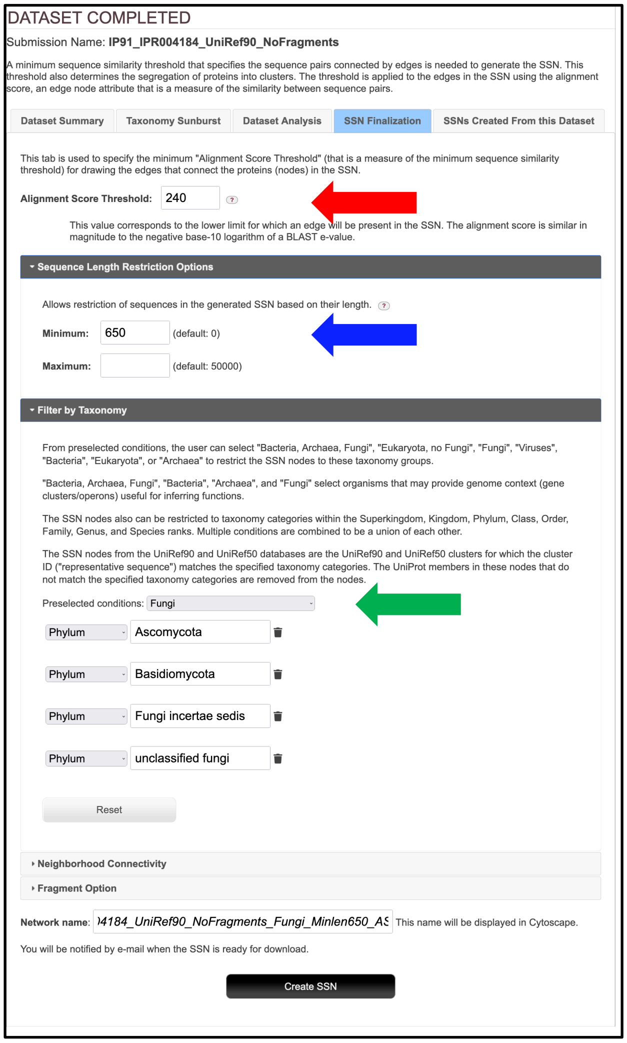
The SSN nodes from the UniRef90 and UniRef50 databases are the UniRef90 and UniRef50 clusters for which the cluster ID ("representative sequence") matches the specified taxonomy categories. The UniProt members in these nodes that do not match the specified taxonomy categories are removed from the nodes.

Preselected conditions: Bacteria

Superkingdom Bacteria

Network name: *UniRef90_NoFragments_Bacteria_Minlen650_AS240* This name will be displayed in Cytoscape.

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



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.

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: 240

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: 650 (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 Ascomycota

Phylum Basidiomycota

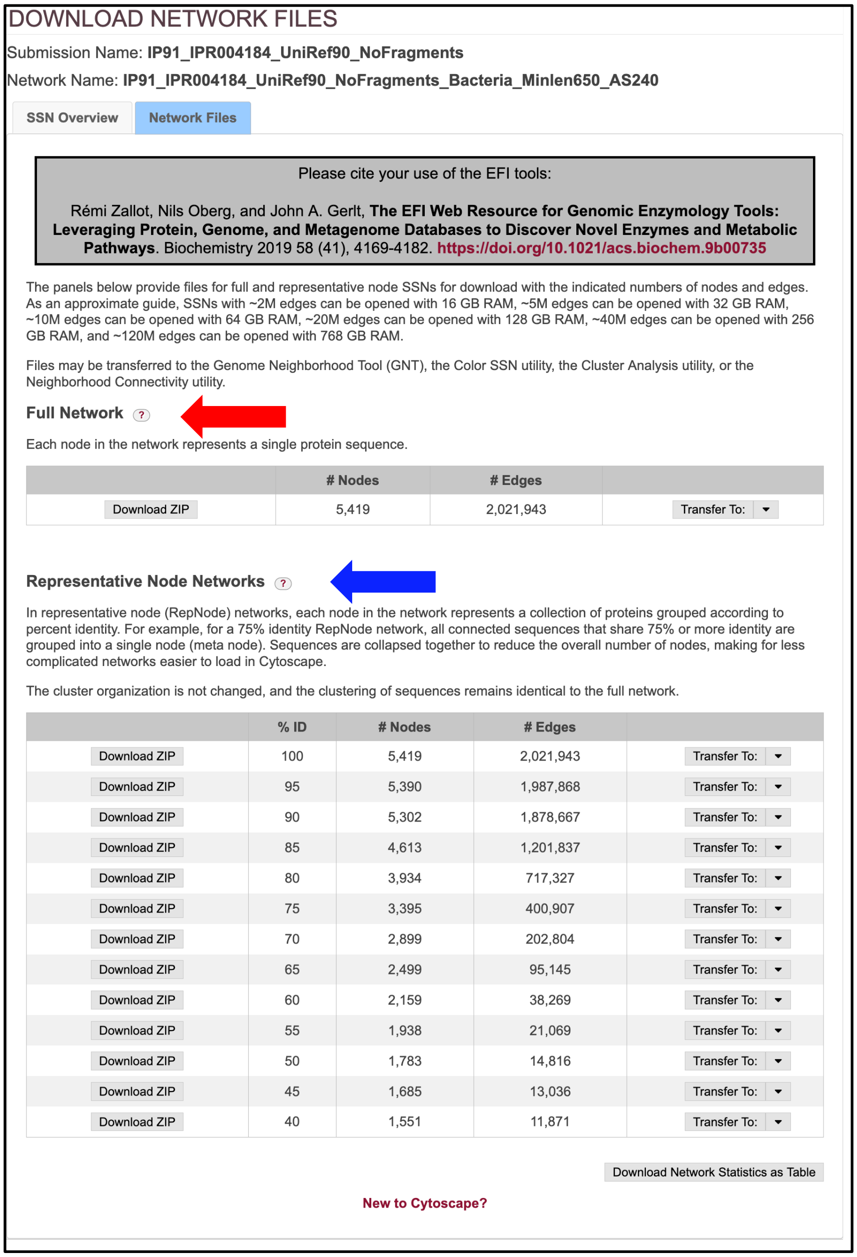
Phylum Fungi incertae sedis

Phylum unclassified fungi

Network name: *4184_UniRef90_NoFragments_Fungi_Minlen650_AS* This name will be 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** 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**

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 ?

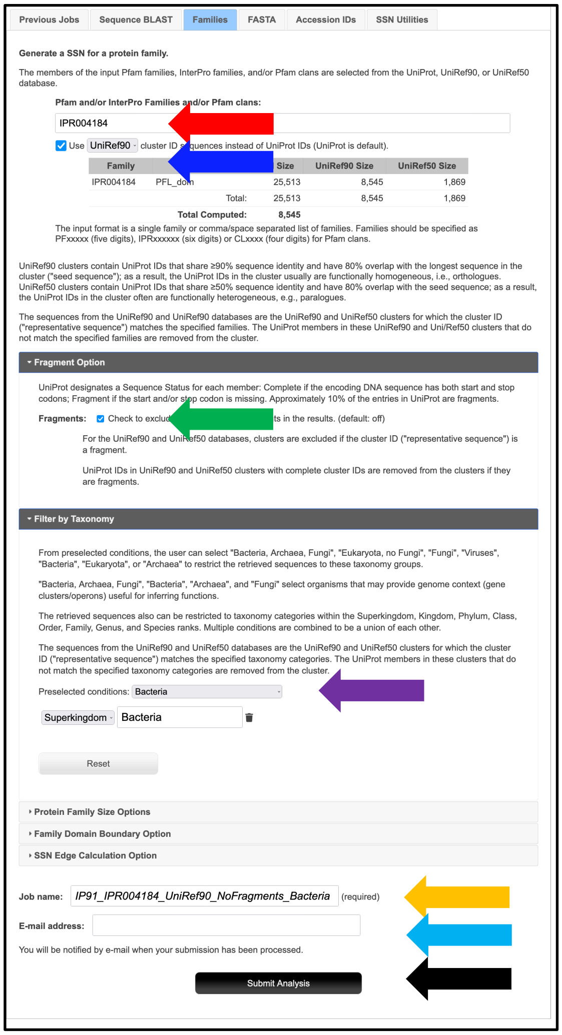
In representative node (RepNode) networks, each node in the network represents a collection of proteins grouped according to percent identity. For example, for a 75% identity RepNode network, all connected sequences that share 75% or more identity are grouped into a single node (meta node). Sequences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

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).



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 s quences 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 UniRef90 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 Uni/Ref50 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 exclud ts 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

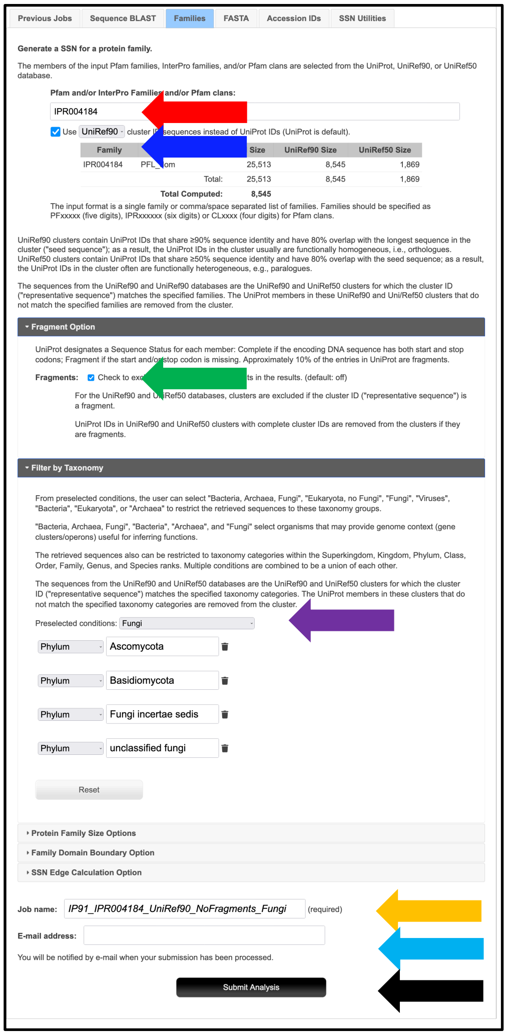
Superkingdom: Bacteria

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.



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 I sequences instead of UniProt IDs (UniProt is default).

Family		Size	UniRef90 Size	UniRef50 Size
IPR004184	PFL_ om	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 UniRef90 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 Uni/Ref50 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 exc ts 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

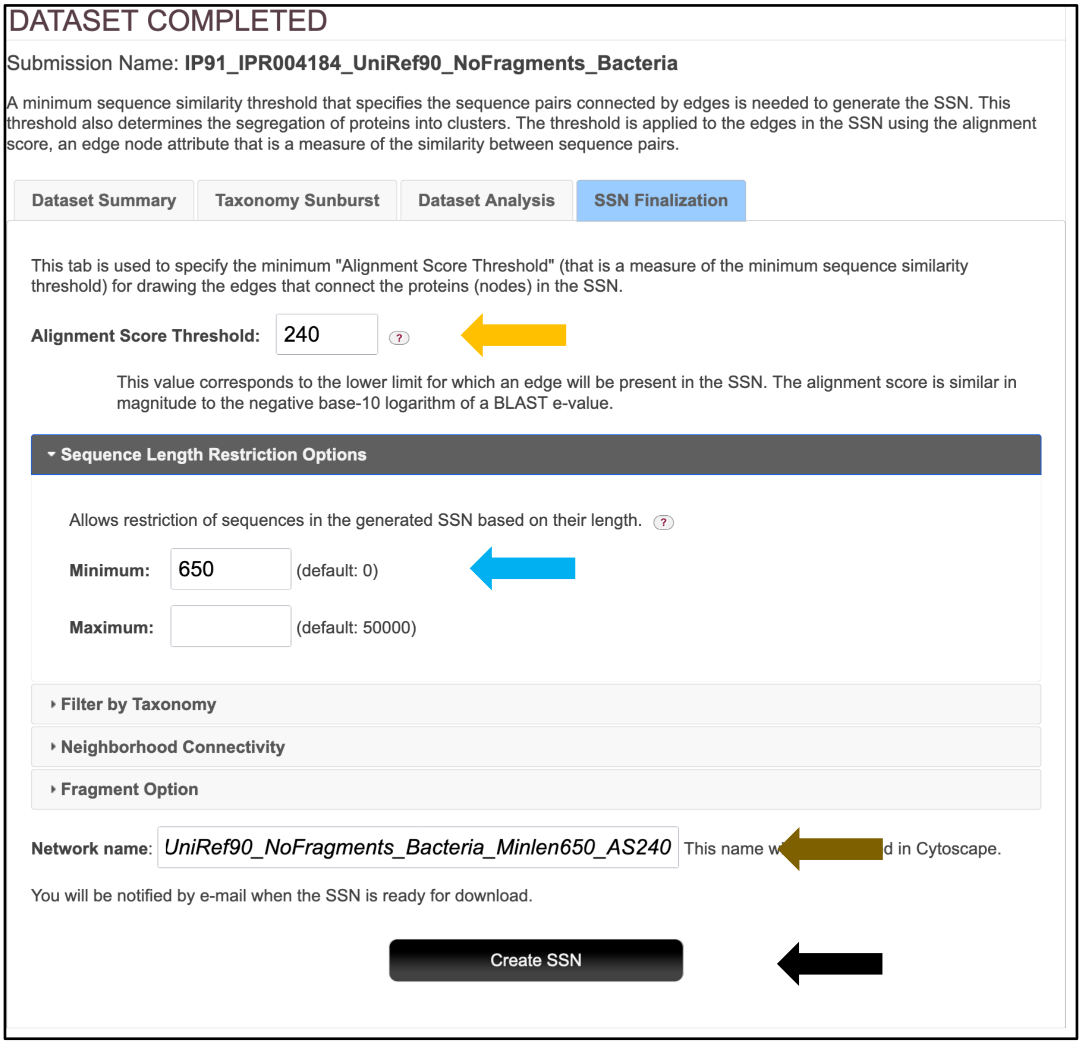
Reset
Protein Family Size Options
Family Domain Boundary Option
SSN Edge Calculation Option

Job name: IP91_IPR004184_UniRef90_NoFragments_Fungi (required)

E-mail address:

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

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.

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: 240

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.

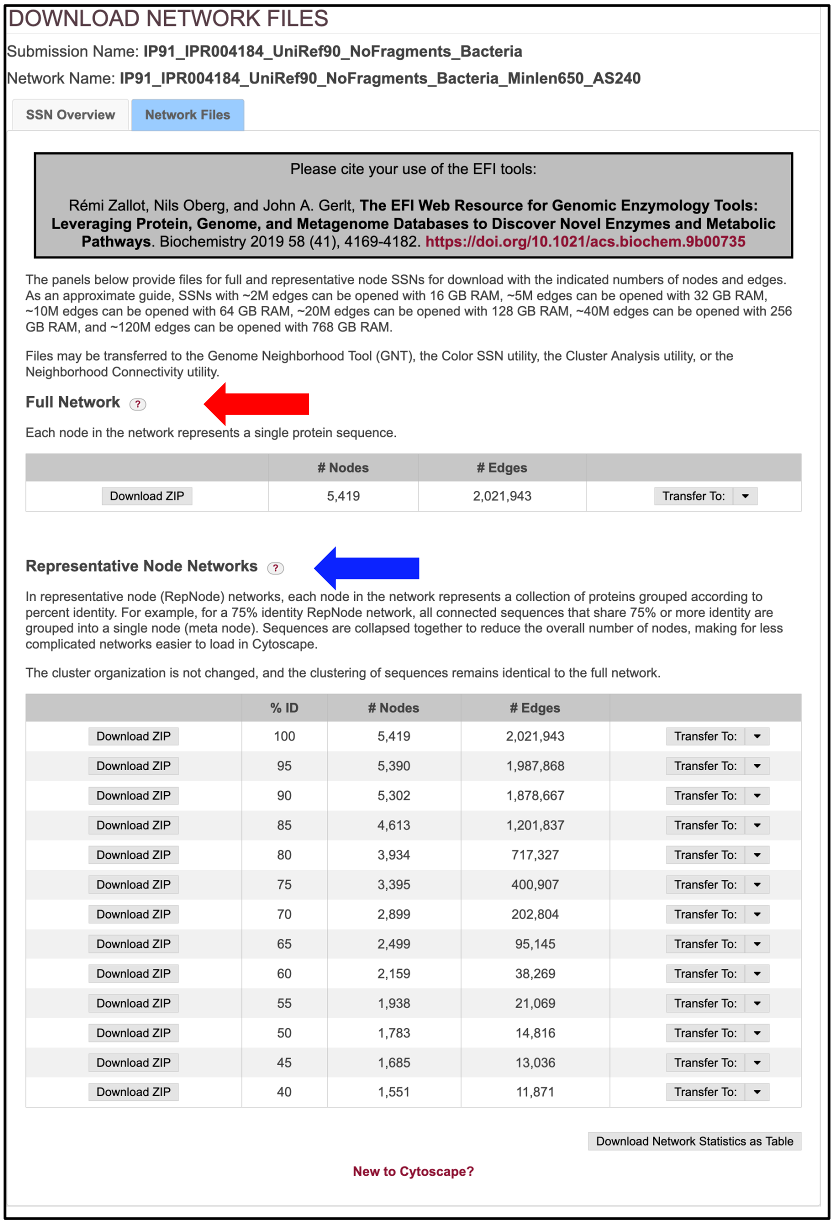
Minimum: 650 (default: 0)

Maximum: (default: 50000)

Network name: *UniRef90_NoFragments_Bacteria_Minlen650_AS240* This name w d 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_UniRef90_NoFragments_Bacteria**

Network Name: **IP91_IPR004184_UniRef90_NoFragments_Bacteria_Minlen650_AS240**

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.

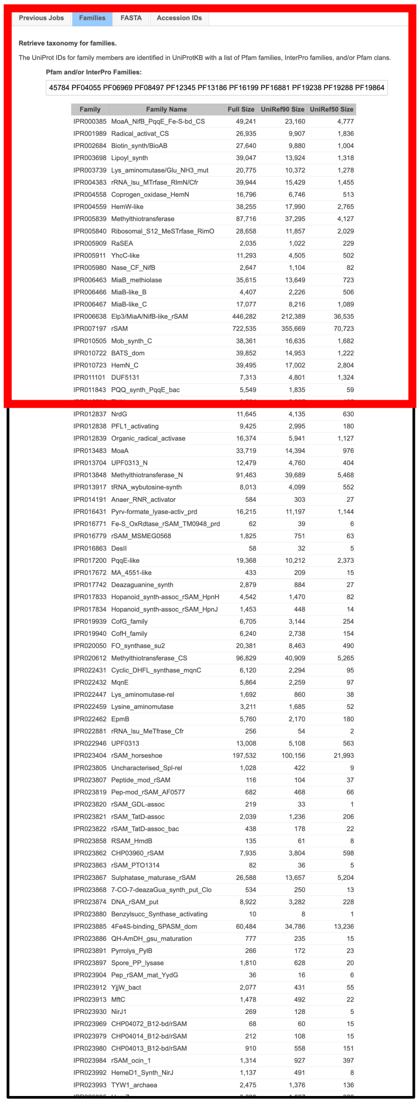
	% ID	# Nodes	# Edges	
Download ZIP	100	5,419	2,021,943	Transfer To: ▾
Download ZIP	95	5,390	1,987,868	Transfer To: ▾
Download ZIP	90	5,302	1,878,667	Transfer To: ▾
Download ZIP	85	4,613	1,201,837	Transfer To: ▾
Download ZIP	80	3,934	717,327	Transfer To: ▾
Download ZIP	75	3,395	400,907	Transfer To: ▾
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: ▾

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

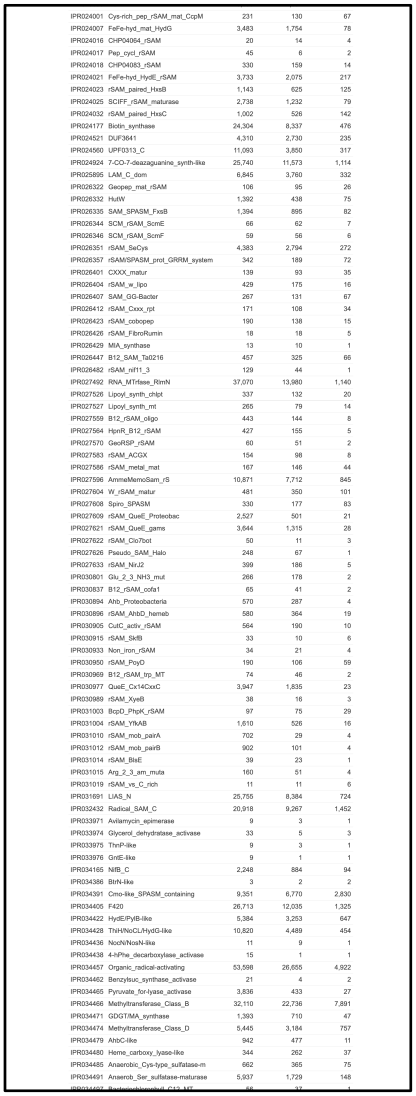
Complete Sequences and Fragments. The **Taxonomy Sunburst** for complete sequences and fragments was generated from UniProt Release 2022_04 by entering a list of 211 Pfam and InterPro families and/or domains (**Tutorial Table 1**) into the **Pfam and/or InterPro Families and/or Pfam Clans** box on the **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.

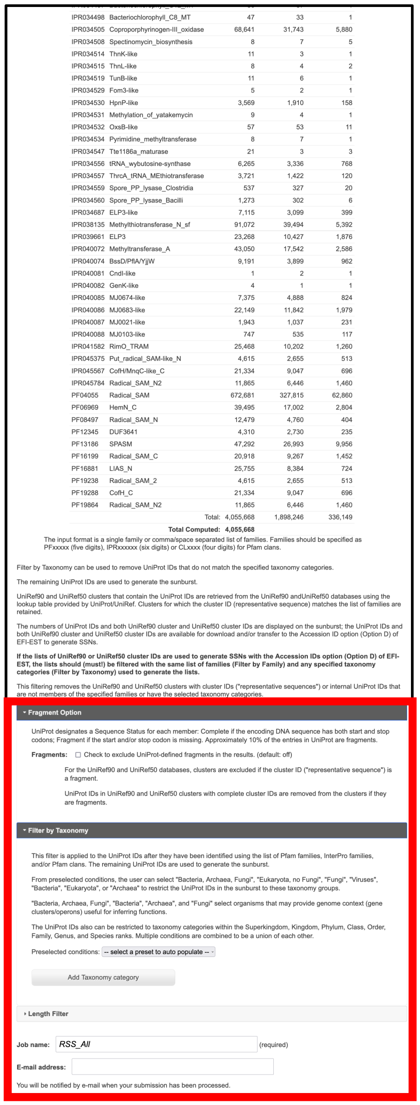


Retrieve taxonomy for families.

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_lsu_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
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	Methylthiotransferase_N	91,463	39,689	5,468
IPR013917	tRNA_wybutosine-synth	8,013	4,099	552
IPR014191	Anaer_RNR_activator	584	303	27
IPR016431	Pyrv-formate_lyase-activ_prd	16,215	11,197	1,144
IPR016771	Fe-S_OxRdtase_rSAM_TM0948_prd	62	39	6
IPR016779	rSAM_MSMEG0568	1,825	751	63
IPR016863	DesII	58	32	5
IPR017200	PqqE-like	19,368	10,212	2,373
IPR017672	MA_4551-like	433	209	15
IPR017742	Deazaguanine_synth	2,879	884	27
IPR017833	Hopanoid_synth-assoc_rSAM_HpnH	4,542	1,470	82
IPR017834	Hopanoid_synth-assoc_rSAM_HpnJ	1,453	448	14
IPR019939	CofG_family	6,705	3,144	254
IPR019940	CofH_family	6,240	2,738	154
IPR020050	FO_synthase_su2	20,381	8,463	490
IPR020612	Methylthiotransferase_CS	96,829	40,909	5,265
IPR022431	Cyclic_DHFL_synthase_mqnC	6,120	2,294	95
IPR022432	MqnE	5,864	2,259	97
IPR022447	Lys_aminomutase-rel	1,692	860	38
IPR022459	Lysine_aminomutase	3,211	1,685	52
IPR022462	EpmB	5,760	2,170	180
IPR022881	rRNA_lsu_MeTfrase_Cfr	256	54	2
IPR022946	UPF0313	13,008	5,108	563
IPR023404	rSAM_horseshoe	197,532	100,156	21,993
IPR023805	Uncharacterised_Spl-rel	1,028	422	9
IPR023807	Peptide_mod_rSAM	116	104	37
IPR023819	Pep-mod_rSAM_AF0577	682	468	66
IPR023820	rSAM_GDL-assoc	219	33	1
IPR023821	rSAM_TatD-assoc	2,039	1,236	206
IPR023822	rSAM_TatD-assoc_bac	438	178	22
IPR023858	RSAM_HmdB	135	61	8
IPR023862	CHP03960_rSAM	7,935	3,804	598
IPR023863	rSAM_PTO1314	82	36	5
IPR023867	Sulphatase_maturase_rSAM	26,588	13,657	5,204
IPR023868	7-CO-7-deazaGua_synth_put_Clo	534	250	13
IPR023874	DNA_rSAM_put	8,922	3,282	228
IPR023880	Benzylsucc_Synthase_activating	10	8	1
IPR023885	4Fe4S-binding_SPASM_dom	60,484	34,786	13,236
IPR023886	QH-AmDH_gsu_maturation	777	235	15
IPR023891	Pyrrolys_PylB	266	172	23
IPR023897	Spore_PP_lysase	1,810	628	20
IPR023904	Pep_rSAM_mat_YydG	36	16	6
IPR023912	YjjW_bact	2,077	431	55
IPR023913	MftC	1,478	492	22
IPR023930	NirJ1	269	128	5
IPR023969	CHP04072_B12-bd/rSAM	68	60	15
IPR023979	CHP04014_B12-bd/rSAM	212	108	15
IPR023980	CHP04013_B12-bd/rSAM	910	558	151
IPR023984	rSAM_ocin_1	1,314	927	397
IPR023992	HemeD1_Synth_NirJ	1,137	491	8
IPR023993	TYW1_archaea	2,475	1,376	136
IPR024001	Cys-rich_pep_rSAM_mat_CcpM	231	130	67
IPR024007	FeFe-hyd_mat_HydG	3,483	1,754	78
IPR024016	CHP04064_rSAM	20	14	4
IPR024017	Pep_cycl_rSAM	45	6	2
IPR024018	CHP04083_rSAM	330	159	14
IPR024021	FeFe-hyd_HydE_rSAM	3,733	2,075	217
IPR024023	rSAM_paired_HxsB	1,143	625	125
IPR024025	SCIFF_rSAM_maturase	2,738	1,232	79
IPR024032	rSAM_paired_HxsC	1,002	526	142
IPR024177	Biotin_synthase	24,304	8,337	476
IPR024521	DUF3641	4,310	2,730	235
IPR024560	UPF0313_C	11,093	3,850	317
IPR024924	7-CO-7-deazaguanine_synth-like	25,740	11,573	1,114
IPR025895	LAM_C_dom	6,845	3,760	332
IPR026322	Geopep_mat_rSAM	106	95	26
IPR026332	HutW	1,392	438	75
IPR026335	SAM_SPASM_FxsB	1,394	895	82
IPR026344	SCM_rSAM_ScmE	66	62	7
IPR026346	SCM_rSAM_ScmF	59	56	6
IPR026351	rSAM_SeCys	4,383	2,794	272
IPR026357	rSAM/SPASM_prot_GRRM_system	342	189	72
IPR026401	CXXX_matur	139	93	35
IPR026404	rSAM_w_lipo	429	175	16
IPR026407	SAM_GG-Bacter	267	131	67
IPR026412	rSAM_Cxxx_rpt	171	108	34
IPR026423	rSAM_cobopep	190	138	15
IPR026426	rSAM_FibroRumin	18	18	5
IPR026429	MIA_synthase	13	10	1
IPR026447	B12_SAM_Ta0216	457	325	66
IPR026482	rSAM_nif11_3	129	44	1
IPR027492	RNA_MTrfase_RlmN	37,070	13,980	1,140
IPR027526	Lipoyl_synth_chlpt	337	132	20
IPR027527	Lipoyl_synth_mt	265	79	14
IPR027559	B12_rSAM_oligo	443	144	8
IPR027564	HpnR_B12_rSAM	427	155	5
IPR027570	GeoRSP_rSAM	60	51	2
IPR027583	rSAM_ACGX	154	98	8
IPR027586	rSAM_metal_mat	167	146	44
IPR027596	AmmeMemoSam_rS	10,871	7,712	845
IPR027604	W_rSAM_matur	481	350	101
IPR027608	Spiro_SPASM	330	177	83
IPR027609	rSAM_QueE_Proteobac	2,527	501	21
IPR027621	rSAM_QueE_gams	3,644	1,315	28
IPR027622	rSAM_Clo7bot	50	11	3
IPR027626	Pseudo_SAM_Halo	248	67	1
IPR027633	rSAM_NirJ2	399	186	5
IPR030801	Glu_2_3_NH3_mut	266	178	2
IPR030837	B12_rSAM_cofa1	65	41	2
IPR030894	Ahb_Proteobacteria	570	287	4
IPR030896	rSAM_AhbD_hemeb	580	364	19
IPR030905	CutC_activ_rSAM	564	190	10
IPR030915	rSAM_SkfB	33	10	6
IPR030933	Non_iron_rSAM	34	21	4
IPR030950	rSAM_PoyD	190	106	59
IPR030969	B12_rSAM_trp_MT	74	46	2
IPR030977	QueE_Cx14CxxC	3,947	1,835	23
IPR030989	rSAM_XyeB	38	16	3
IPR031003	BcpD_PhpK_rSAM	97	75	29
IPR031004	rSAM_YfkAB	1,610	526	16
IPR031010	rSAM_mob_pairA	702	29	4
IPR031012	rSAM_mob_pairB	902	101	4
IPR031014	rSAM_BlsE	39	23	1
IPR031015	Arg_2_3_am_muta	160	51	4
IPR031019	rSAM_vs_C_rich	11	11	6
IPR031691	LIAS_N	25,755	8,384	724
IPR032432	Radical_SAM_C	20,918	9,267	1,452
IPR033971	Avilamycin_epimerase	9	3	1
IPR033974	Glycerol_dehydratase_activase	33	5	3
IPR033975	ThnP-like	9	3	1
IPR033976	GntE-like	9	1	1
IPR034165	NifB_C	2,248	884	94
IPR034386	BtrN-like	3	2	2
IPR034391	Cmo-like_SPASM_containing	9,351	6,770	2,830
IPR034405	F420	26,713	12,035	1,325
IPR034422	HydE/PylB-like	5,384	3,253	647
IPR034428	ThiH/NoCL/HydG-like	10,820	4,489	454
IPR034436	NocN/NosN-like	11	9	1
IPR034438	4-hPhe_decarboxylase_activase	15	1	1
IPR034457	Organic_radical-activating	53,598	26,655	4,922
IPR034462	Benzylsuc_synthase_activase	21	4	2
IPR034465	Pyruvate_for-lyase_activase	3,836	433	27
IPR034466	Methyltransferase_Class_B	32,110	22,736	7,891
IPR034471	GDGT/MA_synthase	1,393	710	47
IPR034474	Methyltransferase_Class_D	5,445	3,184	757
IPR034479	AhbC-like	942	477	11
IPR034480	Heme_carboxy_lyase-like	344	262	37
IPR034485	Anaerobic_Cys-type_sulfatase-m	662	365	75
IPR034491	Anaerob_Ser_sulfatase-maturase	5,937	1,729	148
IPR034497	Bacteriochlorophyll_C12_MT	56	37	1
IPR034498	Bacteriochlorophyll_C8_MT	47	33	1
IPR034505	Coproporphyrinogen-III_oxidase	68,641	31,743	5,880
IPR034508	Spectinomycin_biosynthesis	8	7	5
IPR034514	ThnK-like	11	3	1
IPR034515	ThnL-like	8	4	2
IPR034519	TunB-like	11	6	1
IPR034529	Fom3-like	5	2	1
IPR034530	HpnP-like	3,569	1,910	158
IPR034531	Methylation_of_yatakemycin	9	4	1
IPR034532	OxsB-like	57	53	11
IPR034534	Pyrimidine_methyltransferase	8	7	1
IPR034547	Tte1186a_maturase	21	3	3
IPR034556	tRNA_wybutosine-synthase	6,265	3,336	768
IPR034557	ThrcA_tRNA_MEthiotransferase	3,721	1,422	120
IPR034559	Spore_PP_lysase_Clostridia	537	327	20
IPR034560	Spore_PP_lysase_Bacilli	1,273	302	6
IPR034687	ELP3-like	7,115	3,099	399
IPR038135	Methylthiotransferase_N_sf	91,072	39,494	5,392
IPR039661	ELP3	23,268	10,427	1,876
IPR040072	Methyltransferase_A	43,050	17,542	2,586
IPR040074	BssD/PflA/YjjW	9,191	3,899	962
IPR040081	CndI-like	1	2	1
IPR040082	GenK-like	4	1	1
IPR040085	MJ0674-like	7,375	4,888	824
IPR040086	MJ0683-like	22,149	11,842	1,979
IPR040087	MJ0021-like	1,943	1,037	231
IPR040088	MJ0103-like	747	535	117
IPR041582	RimO_TRAM	25,468	10,202	1,260
IPR045375	Put_radical_SAM-like_N	4,615	2,655	513
IPR045567	CofH/MnqC-like_C	21,334	9,047	696
IPR045784	Radical_SAM_N2	11,865	6,446	1,460
PF04055	Radical_SAM	672,681	327,815	62,860
PF06969	HemN_C	39,495	17,002	2,804
PF08497	Radical_SAM_N	12,479	4,760	404
PF12345	DUF3641	4,310	2,730	235
PF13186	SPASM	47,292	26,993	9,956
PF16199	Radical_SAM_C	20,918	9,267	1,452
PF16881	LIAS_N	25,755	8,384	724
PF19238	Radical_SAM_2	4,615	2,655	513
PF19288	CofH_C	21,334	9,047	696
PF19864	Radical_SAM_N2	11,865	6,446	1,460
	Total:	4,055,668	1,898,246	336,149
	Total Computed:	**4,055,668**		

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 andUniRef50 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

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.

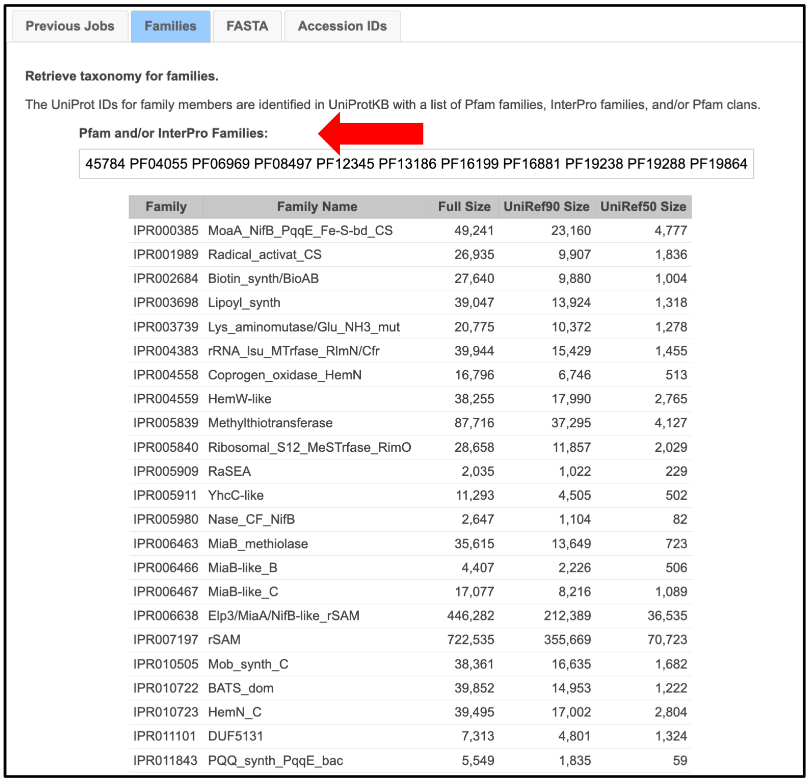
The UniProt IDs also can be restricted to taxonomy categories within the Superkingdom, Kingdom, Phylum, Class, Order, Family, Genus, and Species ranks. Multiple conditions are combined to be a union of each other.

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

Job name: RSS_All (required)

E-mail address:

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

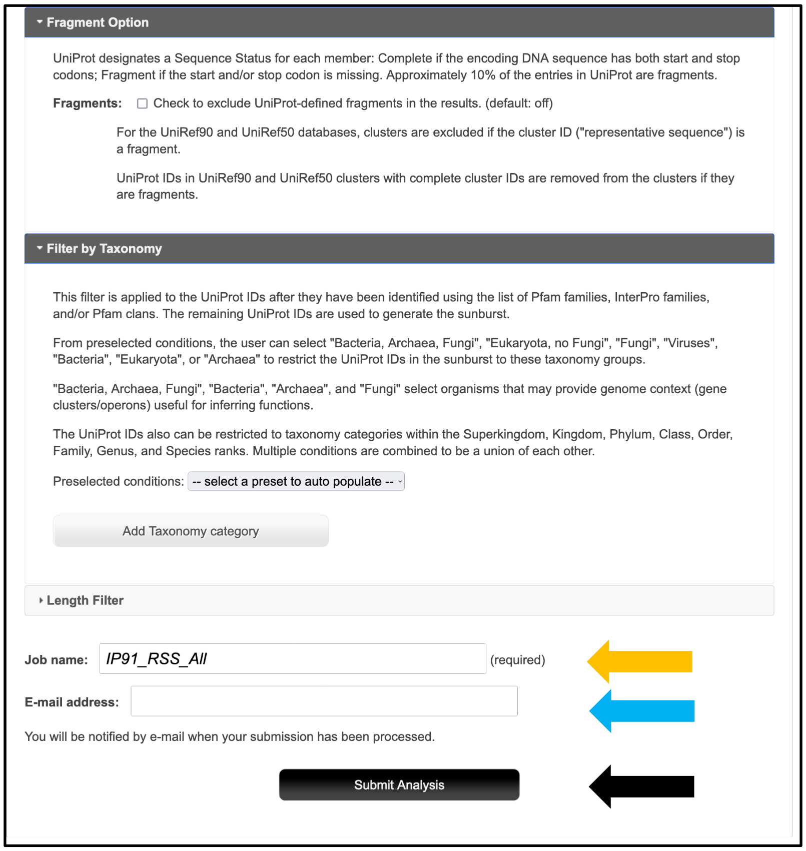


Retrieve taxonomy for families.

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_lsu_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



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.

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.

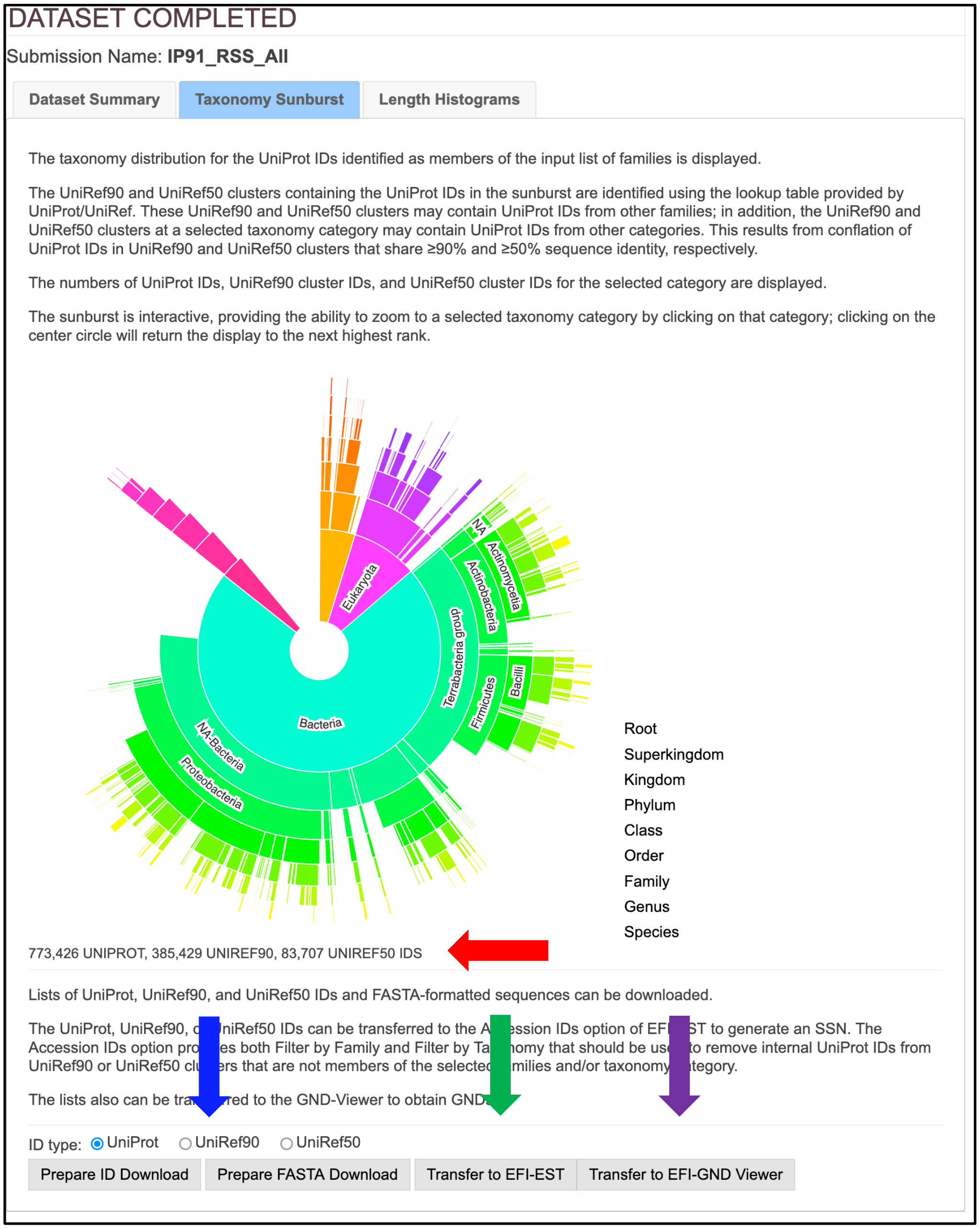
Preselected conditions: -- select a preset to auto populate --

Job name: *IP91_RSS_All* (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**

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 ≥90% and ≥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.

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

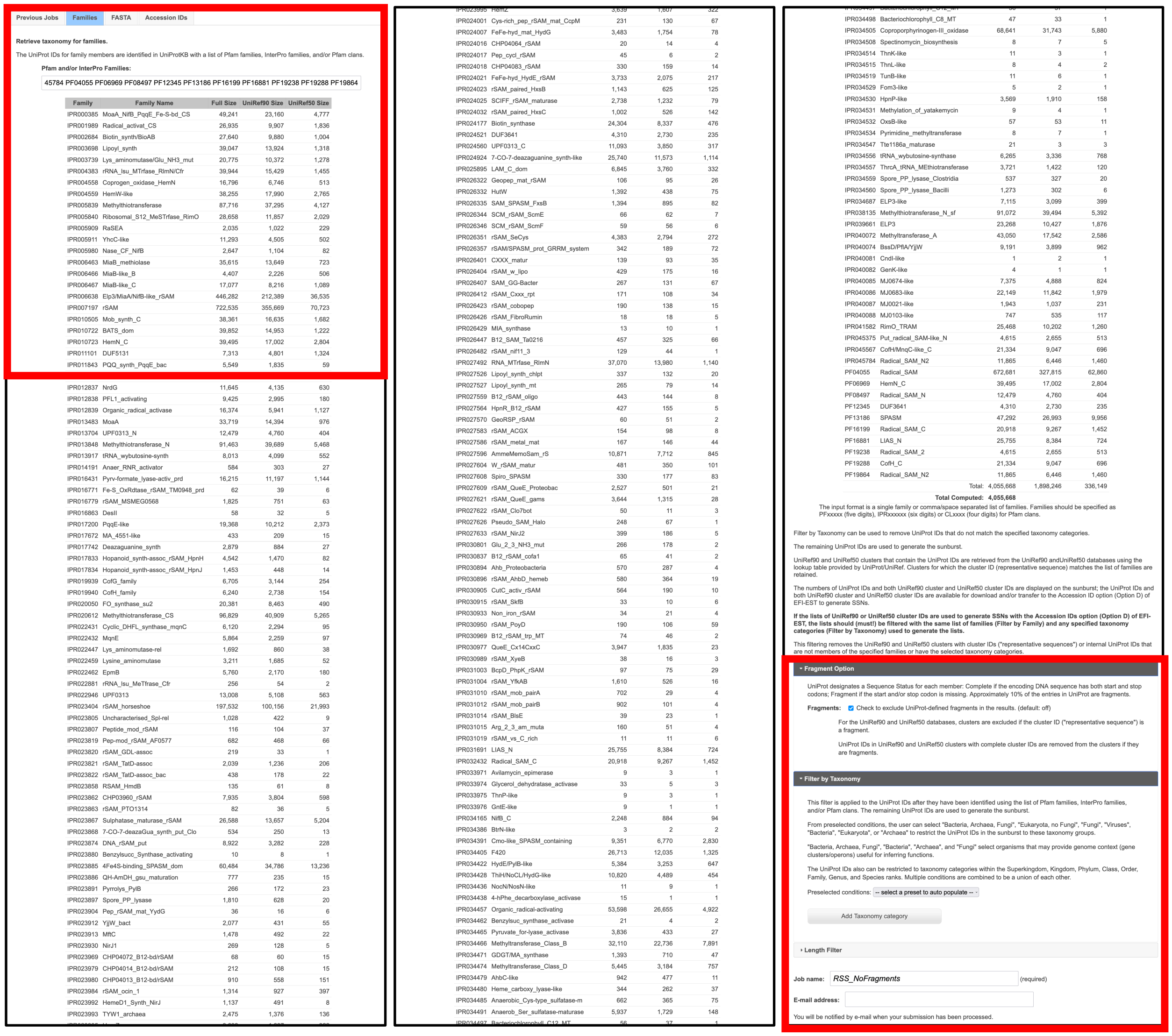
Lists of UniProt, UniRef90, and UniRef50 IDs and FASTA-formatted sequences can be downloaded.

The UniProt, UniRef90, o niRef50 IDs can be transferred to the A ssion IDs option of EF ST to generate an SSN. The Accession IDs option pro es both Filter by Family and Filter by Ta omy that should be us o remove internal UniProt IDs from UniRef90 or UniRef50 cl ers that are not members of the selected milies and/or taxonomy tegory.

The lists also can be tra red to the GND-Viewer to obtain GND

ID type: ● UniProt ○ UniRef90 ○ UniRef50

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.



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_lsu_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
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	Methylthiotransferase_N	91,463	39,689	5,468
IPR013917	tRNA_wybutosine-synth	8,013	4,099	552
IPR014191	Anaer_RNR_activator	584	303	27
IPR016431	Pyrv-formate_lyase-activ_prd	16,215	11,197	1,144
IPR016771	Fe-S_OxRdtase_rSAM_TM0948_prd	62	39	6
IPR016779	rSAM_MSMEG0568	1,825	751	63
IPR016863	DesII	58	32	5
IPR017200	PqqE-like	19,368	10,212	2,373
IPR017672	MA_4551-like	433	209	15
IPR017742	Deazaguanine_synth	2,879	884	27
IPR017833	Hopanoid_synth-assoc_rSAM_HpnH	4,542	1,470	82
IPR017834	Hopanoid_synth-assoc_rSAM_HpnJ	1,453	448	14
IPR019939	CofG_family	6,705	3,144	254
IPR019940	CofH_family	6,240	2,738	154
IPR020050	FO_synthase_su2	20,381	8,463	490
IPR020612	Methylthiotransferase_CS	96,829	40,909	5,265
IPR022431	Cyclic_DHFL_synthase_mqnC	6,120	2,294	95
IPR022432	MqnE	5,864	2,259	97
IPR022447	Lys_aminomutase-rel	1,692	860	38
IPR022459	Lysine_aminomutase	3,211	1,685	52
IPR022462	EpmB	5,760	2,170	180
IPR022881	rRNA_lsu_MeTfrase_Cfr	256	54	2
IPR022946	UPF0313	13,008	5,108	563
IPR023404	rSAM_horseshoe	197,532	100,156	21,993
IPR023805	Uncharacterised_Spl-rel	1,028	422	9
IPR023807	Peptide_mod_rSAM	116	104	37
IPR023819	Pep-mod_rSAM_AF0577	682	468	66
IPR023820	rSAM_GDL-assoc	219	33	1
IPR023821	rSAM_TatD-assoc	2,039	1,236	206
IPR023822	rSAM_TatD-assoc_bac	438	178	22
IPR023858	RSAM_HmdB	135	61	8
IPR023862	CHP03960_rSAM	7,935	3,804	598
IPR023863	rSAM_PTO1314	82	36	5
IPR023867	Sulphatase_maturase_rSAM	26,588	13,657	5,204
IPR023868	7-CO-7-deazaGua_synth_put_Clo	534	250	13
IPR023874	DNA_rSAM_put	8,922	3,282	228
IPR023880	Benzylsucc_Synthase_activating	10	8	1
IPR023885	4Fe4S-binding_SPASM_dom	60,484	34,786	13,236
IPR023886	QH-AmDH_gsu_maturation	777	235	15
IPR023891	Pyrrolys_PylB	266	172	23
IPR023897	Spore_PP_lysase	1,810	628	20
IPR023904	Pep_rSAM_mat_YydG	36	16	6
IPR023912	YjjW_bact	2,077	431	55
IPR023913	MftC	1,478	492	22
IPR023930	NirJ1	269	128	5
IPR023969	CHP04072_B12-bd/rSAM	68	60	15
IPR023979	CHP04014_B12-bd/rSAM	212	108	15
IPR023980	CHP04013_B12-bd/rSAM	910	558	151
IPR023984	rSAM_ocin_1	1,314	927	397
IPR023992	HemeD1_Synth_NirJ	1,137	491	8
IPR023993	TYW1_archaea	2,475	1,376	136
IPR023995	HemZ	3,639	1,607	322
IPR024001	Cys-rich_pep_rSAM_mat_CcpM	231	130	67
IPR024007	FeFe-hyd_mat_HydG	3,483	1,754	78
IPR024016	CHP04064_rSAM	20	14	4
IPR024017	Pep_cycl_rSAM	45	6	2
IPR024018	CHP04083_rSAM	330	159	14
IPR024021	FeFe-hyd_HydE_rSAM	3,733	2,075	217
IPR024023	rSAM_paired_HxsB	1,143	625	125
IPR024025	SCIFF_rSAM_maturase	2,738	1,232	79
IPR024032	rSAM_paired_HxsC	1,002	526	142
IPR024177	Biotin_synthase	24,304	8,337	476
IPR024521	DUF3641	4,310	2,730	235
IPR024560	UPF0313_C	11,093	3,850	317
IPR024924	7-CO-7-deazaguanine_synth-like	25,740	11,573	1,114
IPR025895	LAM_C_dom	6,845	3,760	332
IPR026322	Geopep_mat_rSAM	106	95	26
IPR026332	HutW	1,392	438	75
IPR026335	SAM_SPASM_FxsB	1,394	895	82
IPR026344	SCM_rSAM_ScmE	66	62	7
IPR026346	SCM_rSAM_ScmF	59	56	6
IPR026351	rSAM_SeCys	4,383	2,794	272
IPR026357	rSAM/SPASM_prot_GRRM_system	342	189	72
IPR026401	CXXX_matur	139	93	35
IPR026404	rSAM_w_lipo	429	175	16
IPR026407	SAM_GG-Bacter	267	131	67
IPR026412	rSAM_Cxxx_rpt	171	108	34
IPR026423	rSAM_cobopep	190	138	15
IPR026426	rSAM_FibroRumin	18	18	5
IPR026429	MIA_synthase	13	10	1
IPR026447	B12_SAM_Ta0216	457	325	66
IPR026482	rSAM_nif11_3	129	44	1
IPR027492	RNA_MTrfase_RlmN	37,070	13,980	1,140
IPR027526	Lipoyl_synth_chlpt	337	132	20
IPR027527	Lipoyl_synth_mt	265	79	14
IPR027559	B12_rSAM_oligo	443	144	8
IPR027564	HpnR_B12_rSAM	427	155	5
IPR027570	GeoRSP_rSAM	60	51	2
IPR027583	rSAM_ACGX	154	98	8
IPR027586	rSAM_metal_mat	167	146	44
IPR027596	AmmeMemoSam_rS	10,871	7,712	845
IPR027604	W_rSAM_matur	481	350	101
IPR027608	Spiro_SPASM	330	177	83
IPR027609	rSAM_QueE_Proteobac	2,527	501	21
IPR027621	rSAM_QueE_gams	3,644	1,315	28
IPR027622	rSAM_Clo7bot	50	11	3
IPR027626	Pseudo_SAM_Halo	248	67	1
IPR027633	rSAM_NirJ2	399	186	5
IPR030801	Glu_2_3_NH3_mut	266	178	2
IPR030837	B12_rSAM_cofa1	65	41	2
IPR030894	Ahb_Proteobacteria	570	287	4
IPR030896	rSAM_AhbD_hemeb	580	364	19
IPR030905	CutC_activ_rSAM	564	190	10
IPR030915	rSAM_SkfB	33	10	6
IPR030933	Non_iron_rSAM	34	21	4
IPR030950	rSAM_PoyD	190	106	59
IPR030969	B12_rSAM_trp_MT	74	46	2
IPR030977	QueE_Cx14CxxC	3,947	1,835	23
IPR030989	rSAM_XyeB	38	16	3
IPR031003	BcpD_PhpK_rSAM	97	75	29
IPR031004	rSAM_YfkAB	1,610	526	16
IPR031010	rSAM_mob_pairA	702	29	4
IPR031012	rSAM_mob_pairB	902	101	4
IPR031014	rSAM_BlsE	39	23	1
IPR031015	Arg_2_3_am_muta	160	51	4
IPR031019	rSAM_vs_C_rich	11	11	6
IPR031691	LIAS_N	25,755	8,384	724
IPR032432	Radical_SAM_C	20,918	9,267	1,452
IPR033971	Avilamycin_epimerase	9	3	1
IPR033974	Glycerol_dehydratase_activase	33	5	3
IPR033975	ThnP-like	9	3	1
IPR033976	GntE-like	9	1	1
IPR034165	NifB_C	2,248	884	94
IPR034386	BtrN-like	3	2	2
IPR034391	Cmo-like_SPASM_containing	9,351	6,770	2,830
IPR034405	F420	26,713	12,035	1,325
IPR034422	HydE/PylB-like	5,384	3,253	647
IPR034428	ThiH/NoCL/HydG-like	10,820	4,489	454
IPR034436	NocN/NosN-like	11	9	1
IPR034438	4-hPhe_decarboxylase_activase	15	1	1
IPR034457	Organic_radical-activating	53,598	26,655	4,922
IPR034462	Benzylsuc_synthase_activase	21	4	2
IPR034465	Pyruvate_for-lyase_activase	3,836	433	27
IPR034466	Methyltransferase_Class_B	32,110	22,736	7,891
IPR034471	GDGT/MA_synthase	1,393	710	47
IPR034474	Methyltransferase_Class_D	5,445	3,184	757
IPR034479	AhbC-like	942	477	11
IPR034480	Heme_carboxy_lyase-like	344	262	37
IPR034485	Anaerobic_Cys-type_sulfatase-m	662	365	75
IPR034491	Anaerob_Ser_sulfatase-maturase	5,937	1,729	148
IPR034497	Bacteriochlorophyll_C12_MT	56	37	1
IPR034498	Bacteriochlorophyll_C8_MT	47	33	1
IPR034505	Coproporphyrinogen-III_oxidase	68,641	31,743	5,880
IPR034508	Spectinomycin_biosynthesis	8	7	5
IPR034514	ThnK-like	11	3	1
IPR034515	ThnL-like	8	4	2
IPR034519	TunB-like	11	6	1
IPR034529	Fom3-like	5	2	1
IPR034530	HpnP-like	3,569	1,910	158
IPR034531	Methylation_of_yatakemycin	9	4	1
IPR034532	OxsB-like	57	53	11
IPR034534	Pyrimidine_methyltransferase	8	7	1
IPR034547	Tte1186a_maturase	21	3	3
IPR034556	tRNA_wybutosine-synthase	6,265	3,336	768
IPR034557	ThrcA_tRNA_MEthiotransferase	3,721	1,422	120
IPR034559	Spore_PP_lysase_Clostridia	537	327	20
IPR034560	Spore_PP_lysase_Bacilli	1,273	302	6
IPR034687	ELP3-like	7,115	3,099	399
IPR038135	Methylthiotransferase_N_sf	91,072	39,494	5,392
IPR039661	ELP3	23,268	10,427	1,876
IPR040072	Methyltransferase_A	43,050	17,542	2,586
IPR040074	BssD/PflA/YjjW	9,191	3,899	962
IPR040081	CndI-like	1	2	1
IPR040082	GenK-like	4	1	1
IPR040085	MJ0674-like	7,375	4,888	824
IPR040086	MJ0683-like	22,149	11,842	1,979
IPR040087	MJ0021-like	1,943	1,037	231
IPR040088	MJ0103-like	747	535	117
IPR041582	RimO_TRAM	25,468	10,202	1,260
IPR045375	Put_radical_SAM-like_N	4,615	2,655	513
IPR045567	CofH/MnqC-like_C	21,334	9,047	696
IPR045784	Radical_SAM_N2	11,865	6,446	1,460
PF04055	Radical_SAM	672,681	327,815	62,860
PF06969	HemN_C	39,495	17,002	2,804
PF08497	Radical_SAM_N	12,479	4,760	404
PF12345	DUF3641	4,310	2,730	235
PF13186	SPASM	47,292	26,993	9,956
PF16199	Radical_SAM_C	20,918	9,267	1,452
PF16881	LIAS_N	25,755	8,384	724
PF19238	Radical_SAM_2	4,615	2,655	513
PF19288	CofH_C	21,334	9,047	696
PF19864	Radical_SAM_N2	11,865	6,446	1,460
	Total:	4,055,668	1,898,246	336,149
	Total Computed:	**4,055,668**		

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 andUniRef50 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

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.

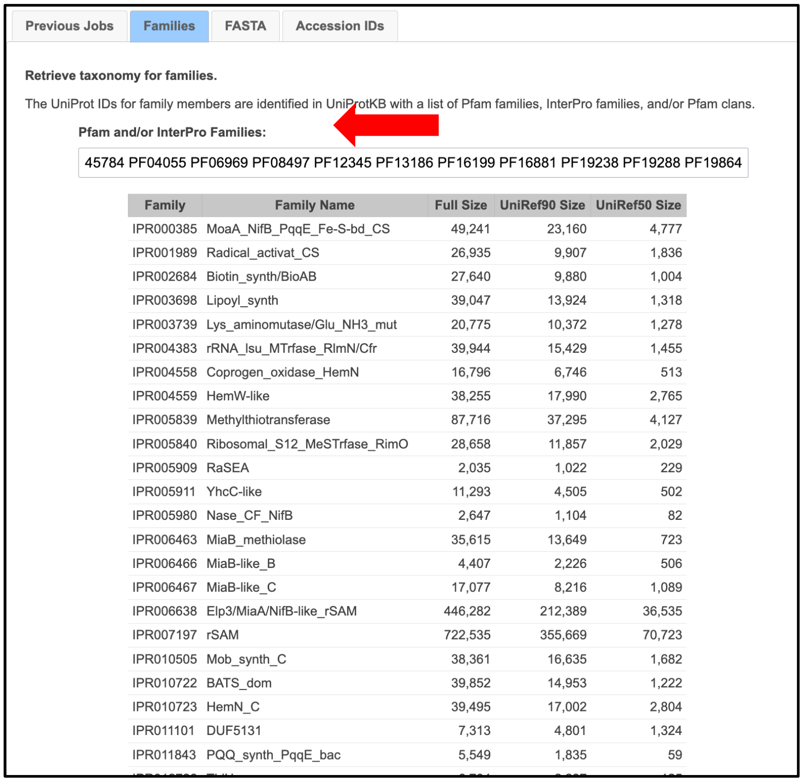
The UniProt IDs also can be restricted to taxonomy categories within the Superkingdom, Kingdom, Phylum, Class, Order, Family, Genus, and Species ranks. Multiple conditions are combined to be a union of each other.

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

Job name: RSS_NoFragments (required)

E-mail address:

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



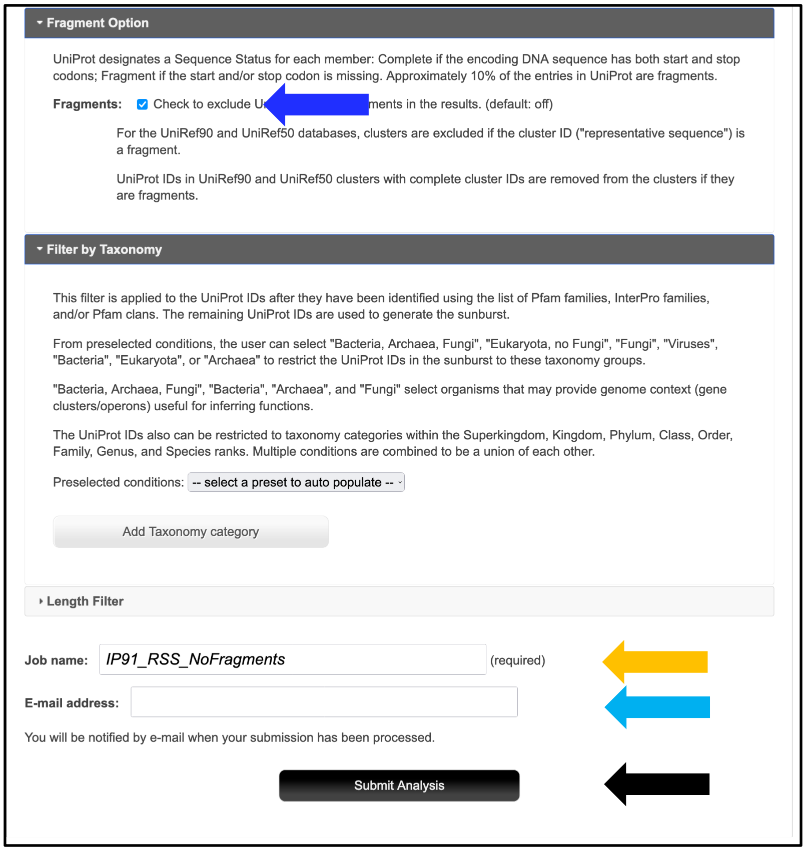
Retrieve taxonomy for families.

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_lsu_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



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.

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.

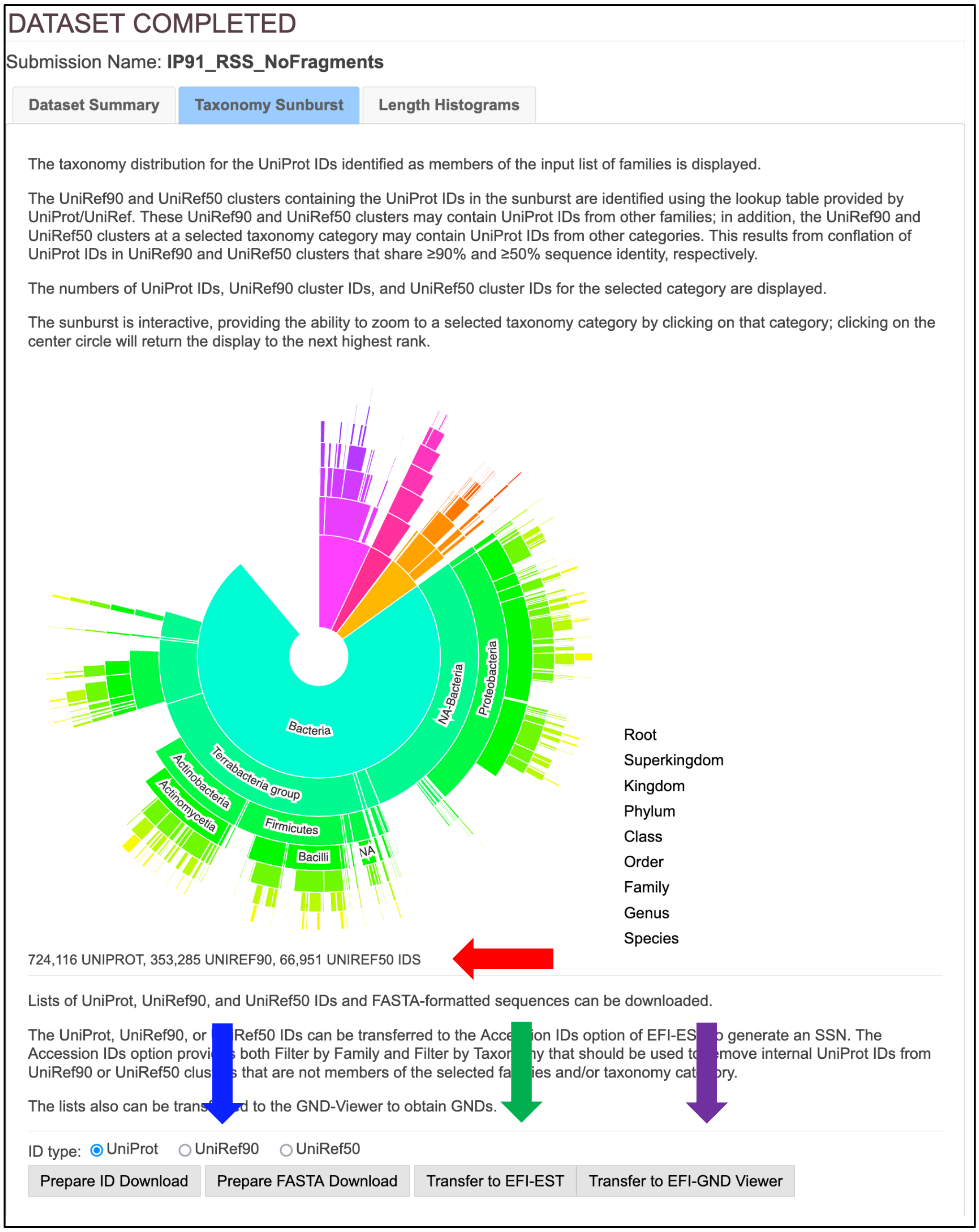
Preselected conditions: -- select a preset to auto populate --

Job name: *IP91_RSS_NoFragments* (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_NoFragments**

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 ≥90% and ≥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.

724,116 UNIPROT, 353,285 UNIREF90, 66,951 UNIREF50 IDS

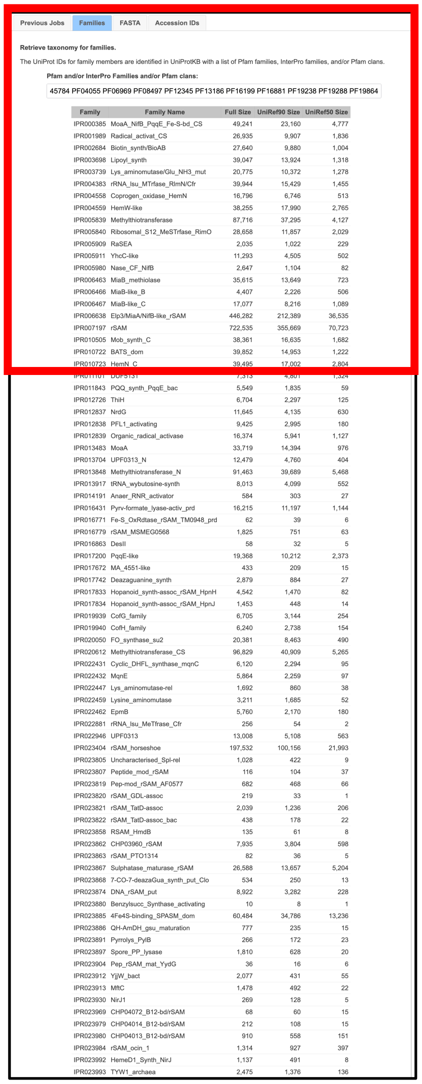
Lists of UniProt, UniRef90, and UniRef50 IDs and FASTA-formatted sequences can be downloaded.

The UniProt, UniRef90, or Ref50 IDs can be transferred to the Acce ion IDs option of EFI-ES o generate an SSN. The Accession IDs option provi s both Filter by Family and Filter by Taxon ny that should be used t move internal UniProt IDs from UniRef90 or UniRef50 clus s that are not members of the selected fa ies and/or taxonomy cat ory.

The lists also can be trans d to the GND-Viewer to obtain GNDs.

ID type: ● UniProt ○ UniRef90 ○ UniRef50

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

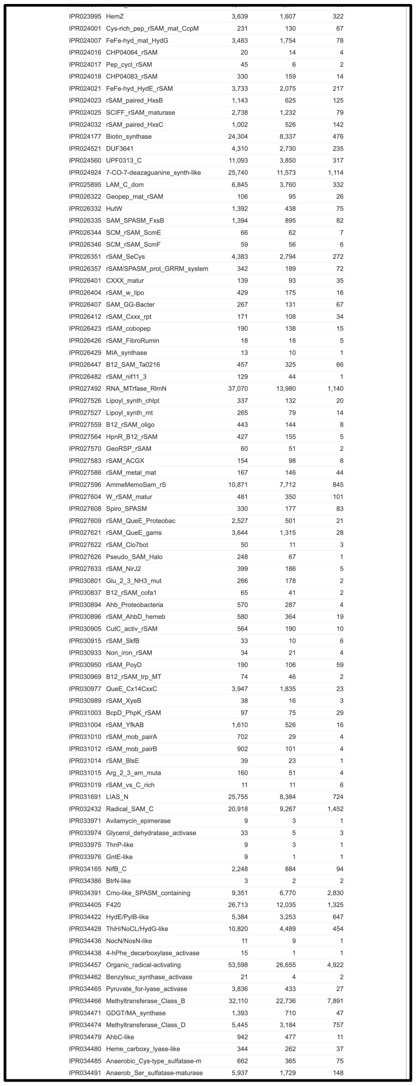
Retrieve taxonomy for families.

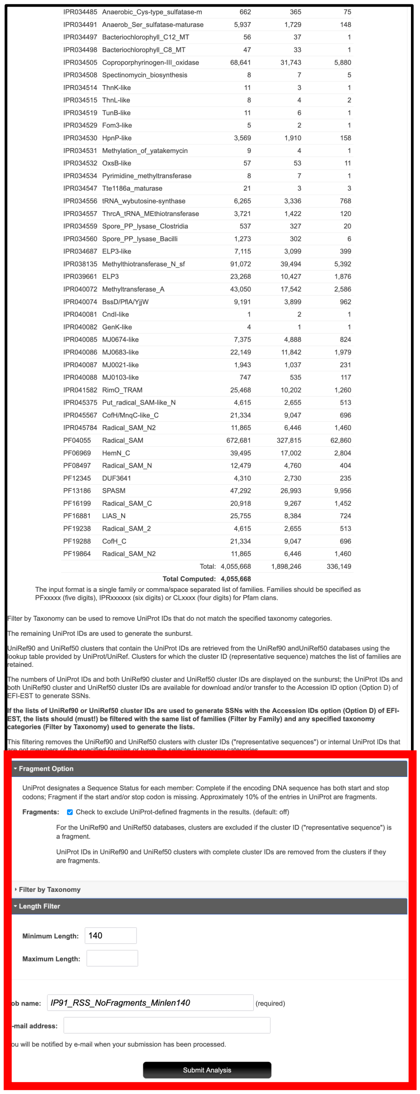
The UniProt IDs for family members are identified in UniProtKB with a list of Pfam families, InterPro families, and/or Pfam clans.

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_lsu_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
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	Methylthiotransferase_N	91,463	39,689	5,468
IPR013917	tRNA_wybutosine-synth	8,013	4,099	552
IPR014191	Anaer_RNR_activator	584	303	27
IPR016431	Pyrv-formate_lyase-activ_prd	16,215	11,197	1,144
IPR016771	Fe-S_OxRdtase_rSAM_TM0948_prd	62	39	6
IPR016779	rSAM_MSMEG0568	1,825	751	63
IPR016863	DesII	58	32	5
IPR017200	PqqE-like	19,368	10,212	2,373
IPR017672	MA_4551-like	433	209	15
IPR017742	Deazaguanine_synth	2,879	884	27
IPR017833	Hopanoid_synth-assoc_rSAM_HpnH	4,542	1,470	82
IPR017834	Hopanoid_synth-assoc_rSAM_HpnJ	1,453	448	14
IPR019939	CofG_family	6,705	3,144	254
IPR019940	CofH_family	6,240	2,738	154
IPR020050	FO_synthase_su2	20,381	8,463	490
IPR020612	Methylthiotransferase_CS	96,829	40,909	5,265
IPR022431	Cyclic_DHFL_synthase_mqnC	6,120	2,294	95
IPR022432	MqnE	5,864	2,259	97
IPR022447	Lys_aminomutase-rel	1,692	860	38
IPR022459	Lysine_aminomutase	3,211	1,685	52
IPR022462	EpmB	5,760	2,170	180
IPR022881	rRNA_lsu_MeTfrase_Cfr	256	54	2
IPR022946	UPF0313	13,008	5,108	563
IPR023404	rSAM_horseshoe	197,532	100,156	21,993
IPR023805	Uncharacterised_Spl-rel	1,028	422	9
IPR023807	Peptide_mod_rSAM	116	104	37
IPR023819	Pep-mod_rSAM_AF0577	682	468	66
IPR023820	rSAM_GDL-assoc	219	33	1
IPR023821	rSAM_TatD-assoc	2,039	1,236	206
IPR023822	rSAM_TatD-assoc_bac	438	178	22
IPR023858	RSAM_HmdB	135	61	8
IPR023862	CHP03960_rSAM	7,935	3,804	598
IPR023863	rSAM_PTO1314	82	36	5
IPR023867	Sulphatase_maturase_rSAM	26,588	13,657	5,204
IPR023868	7-CO-7-deazaGua_synth_put_Clo	534	250	13
IPR023874	DNA_rSAM_put	8,922	3,282	228
IPR023880	Benzylsucc_Synthase_activating	10	8	1
IPR023885	4Fe4S-binding_SPASM_dom	60,484	34,786	13,236
IPR023886	QH-AmDH_gsu_maturation	777	235	15
IPR023891	Pyrrolys_PylB	266	172	23
IPR023897	Spore_PP_lysase	1,810	628	20
IPR023904	Pep_rSAM_mat_YydG	36	16	6
IPR023912	YjjW_bact	2,077	431	55
IPR023913	MftC	1,478	492	22
IPR023930	NirJ1	269	128	5
IPR023969	CHP04072_B12-bd/rSAM	68	60	15
IPR023979	CHP04014_B12-bd/rSAM	212	108	15
IPR023980	CHP04013_B12-bd/rSAM	910	558	151
IPR023984	rSAM_ocin_1	1,314	927	397
IPR023992	HemeD1_Synth_NirJ	1,137	491	8
IPR023993	TYW1_archaea	2,475	1,376	136
IPR023995	HemZ	3,639	1,607	322
IPR024001	Cys-rich_pep_rSAM_mat_CcpM	231	130	67
IPR024007	FeFe-hyd_mat_HydG	3,483	1,754	78
IPR024016	CHP04064_rSAM	20	14	4
IPR024017	Pep_cycl_rSAM	45	6	2
IPR024018	CHP04083_rSAM	330	159	14
IPR024021	FeFe-hyd_HydE_rSAM	3,733	2,075	217
IPR024023	rSAM_paired_HxsB	1,143	625	125
IPR024025	SCIFF_rSAM_maturase	2,738	1,232	79
IPR024032	rSAM_paired_HxsC	1,002	526	142
IPR024177	Biotin_synthase	24,304	8,337	476
IPR024521	DUF3641	4,310	2,730	235
IPR024560	UPF0313_C	11,093	3,850	317
IPR024924	7-CO-7-deazaguanine_synth-like	25,740	11,573	1,114
IPR025895	LAM_C_dom	6,845	3,760	332
IPR026322	Geopep_mat_rSAM	106	95	26
IPR026332	HutW	1,392	438	75
IPR026335	SAM_SPASM_FxsB	1,394	895	82
IPR026344	SCM_rSAM_ScmE	66	62	7
IPR026346	SCM_rSAM_ScmF	59	56	6
IPR026351	rSAM_SeCys	4,383	2,794	272
IPR026357	rSAM/SPASM_prot_GRRM_system	342	189	72
IPR026401	CXXX_matur	139	93	35
IPR026404	rSAM_w_lipo	429	175	16
IPR026407	SAM_GG-Bacter	267	131	67
IPR026412	rSAM_Cxxx_rpt	171	108	34
IPR026423	rSAM_cobopep	190	138	15
IPR026426	rSAM_FibroRumin	18	18	5
IPR026429	MIA_synthase	13	10	1
IPR026447	B12_SAM_Ta0216	457	325	66
IPR026482	rSAM_nif11_3	129	44	1
IPR027492	RNA_MTrfase_RlmN	37,070	13,980	1,140
IPR027526	Lipoyl_synth_chlpt	337	132	20
IPR027527	Lipoyl_synth_mt	265	79	14
IPR027559	B12_rSAM_oligo	443	144	8
IPR027564	HpnR_B12_rSAM	427	155	5
IPR027570	GeoRSP_rSAM	60	51	2
IPR027583	rSAM_ACGX	154	98	8
IPR027586	rSAM_metal_mat	167	146	44
IPR027596	AmmeMemoSam_rS	10,871	7,712	845
IPR027604	W_rSAM_matur	481	350	101
IPR027608	Spiro_SPASM	330	177	83
IPR027609	rSAM_QueE_Proteobac	2,527	501	21
IPR027621	rSAM_QueE_gams	3,644	1,315	28
IPR027622	rSAM_Clo7bot	50	11	3
IPR027626	Pseudo_SAM_Halo	248	67	1
IPR027633	rSAM_NirJ2	399	186	5
IPR030801	Glu_2_3_NH3_mut	266	178	2
IPR030837	B12_rSAM_cofa1	65	41	2
IPR030894	Ahb_Proteobacteria	570	287	4
IPR030896	rSAM_AhbD_hemeb	580	364	19
IPR030905	CutC_activ_rSAM	564	190	10
IPR030915	rSAM_SkfB	33	10	6
IPR030933	Non_iron_rSAM	34	21	4
IPR030950	rSAM_PoyD	190	106	59
IPR030969	B12_rSAM_trp_MT	74	46	2
IPR030977	QueE_Cx14CxxC	3,947	1,835	23
IPR030989	rSAM_XyeB	38	16	3
IPR031003	BcpD_PhpK_rSAM	97	75	29
IPR031004	rSAM_YfkAB	1,610	526	16
IPR031010	rSAM_mob_pairA	702	29	4
IPR031012	rSAM_mob_pairB	902	101	4
IPR031014	rSAM_BlsE	39	23	1
IPR031015	Arg_2_3_am_muta	160	51	4
IPR031019	rSAM_vs_C_rich	11	11	6
IPR031691	LIAS_N	25,755	8,384	724
IPR032432	Radical_SAM_C	20,918	9,267	1,452
IPR033971	Avilamycin_epimerase	9	3	1
IPR033974	Glycerol_dehydratase_activase	33	5	3
IPR033975	ThnP-like	9	3	1
IPR033976	GntE-like	9	1	1
IPR034165	NifB_C	2,248	884	94
IPR034386	BtrN-like	3	2	2
IPR034391	Cmo-like_SPASM_containing	9,351	6,770	2,830
IPR034405	F420	26,713	12,035	1,325
IPR034422	HydE/PylB-like	5,384	3,253	647
IPR034428	ThiH/NoCL/HydG-like	10,820	4,489	454
IPR034436	NocN/NosN-like	11	9	1
IPR034438	4-hPhe_decarboxylase_activase	15	1	1
IPR034457	Organic_radical-activating	53,598	26,655	4,922
IPR034462	Benzylsuc_synthase_activase	21	4	2
IPR034465	Pyruvate_for-lyase_activase	3,836	433	27
IPR034466	Methyltransferase_Class_B	32,110	22,736	7,891
IPR034471	GDGT/MA_synthase	1,393	710	47
IPR034474	Methyltransferase_Class_D	5,445	3,184	757
IPR034479	AhbC-like	942	477	11
IPR034480	Heme_carboxy_lyase-like	344	262	37
IPR034485	Anaerobic_Cys-type_sulfatase-m	662	365	75
IPR034491	Anaerob_Ser_sulfatase-maturase	5,937	1,729	148
IPR034497	Bacteriochlorophyll_C12_MT	56	37	1
IPR034498	Bacteriochlorophyll_C8_MT	47	33	1
IPR034505	Coproporphyrinogen-III_oxidase	68,641	31,743	5,880
IPR034508	Spectinomycin_biosynthesis	8	7	5
IPR034514	ThnK-like	11	3	1
IPR034515	ThnL-like	8	4	2
IPR034519	TunB-like	11	6	1
IPR034529	Fom3-like	5	2	1
IPR034530	HpnP-like	3,569	1,910	158
IPR034531	Methylation_of_yatakemycin	9	4	1
IPR034532	OxsB-like	57	53	11
IPR034534	Pyrimidine_methyltransferase	8	7	1
IPR034547	Tte1186a_maturase	21	3	3
IPR034556	tRNA_wybutosine-synthase	6,265	3,336	768
IPR034557	ThrcA_tRNA_MEthiotransferase	3,721	1,422	120
IPR034559	Spore_PP_lysase_Clostridia	537	327	20
IPR034560	Spore_PP_lysase_Bacilli	1,273	302	6
IPR034687	ELP3-like	7,115	3,099	399
IPR038135	Methylthiotransferase_N_sf	91,072	39,494	5,392
IPR039661	ELP3	23,268	10,427	1,876
IPR040072	Methyltransferase_A	43,050	17,542	2,586
IPR040074	BssD/PflA/YjjW	9,191	3,899	962
IPR040081	CndI-like	1	2	1
IPR040082	GenK-like	4	1	1
IPR040085	MJ0674-like	7,375	4,888	824
IPR040086	MJ0683-like	22,149	11,842	1,979
IPR040087	MJ0021-like	1,943	1,037	231
IPR040088	MJ0103-like	747	535	117
IPR041582	RimO_TRAM	25,468	10,202	1,260
IPR045375	Put_radical_SAM-like_N	4,615	2,655	513
IPR045567	CofH/MnqC-like_C	21,334	9,047	696
IPR045784	Radical_SAM_N2	11,865	6,446	1,460
PF04055	Radical_SAM	672,681	327,815	62,860
PF06969	HemN_C	39,495	17,002	2,804
PF08497	Radical_SAM_N	12,479	4,760	404
PF12345	DUF3641	4,310	2,730	235
PF13186	SPASM	47,292	26,993	9,956
PF16199	Radical_SAM_C	20,918	9,267	1,452
PF16881	LIAS_N	25,755	8,384	724
PF19238	Radical_SAM_2	4,615	2,655	513
PF19288	CofH_C	21,334	9,047	696
PF19864	Radical_SAM_N2	11,865	6,446	1,460
	Total:	4,055,668	1,898,246	336,149
	Total Computed:	**4,055,668**		





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 andUniRef50 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

Minimum Length: 140

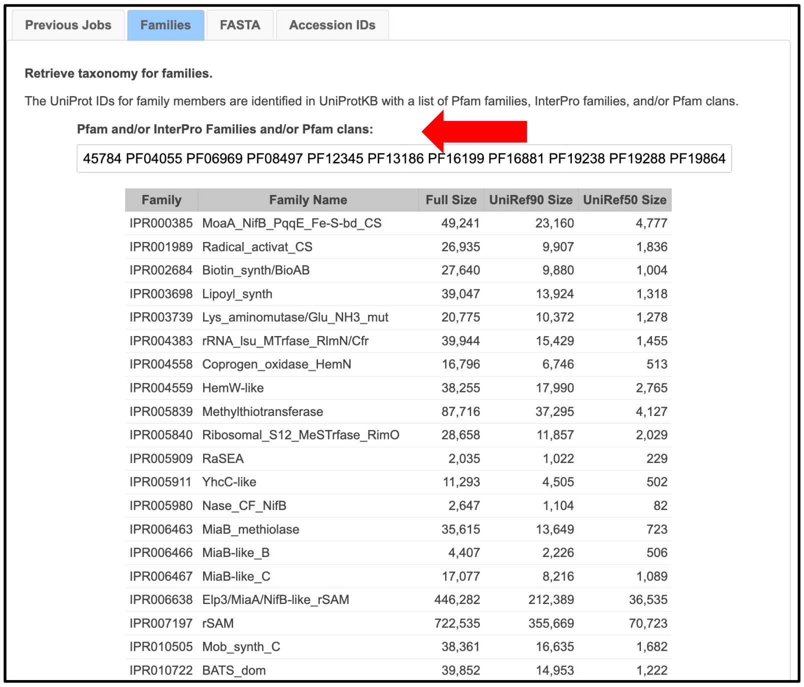
Maximum Length:

ob name: *IP91_RSS_NoFragments_Minlen140* (required)

-mail address:

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

Submit Analysis



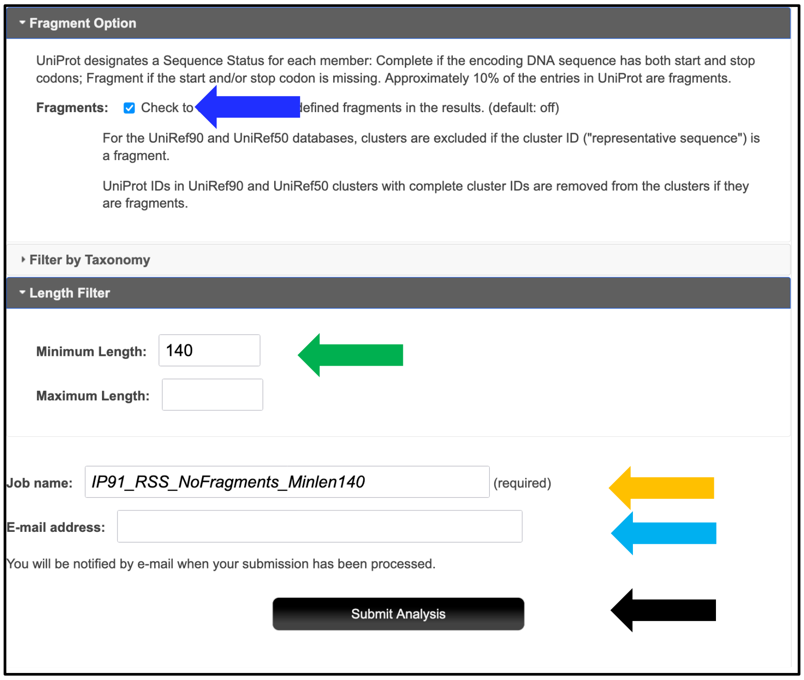
Retrieve taxonomy for families.

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_lsu_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



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 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

Minimum Length: 140

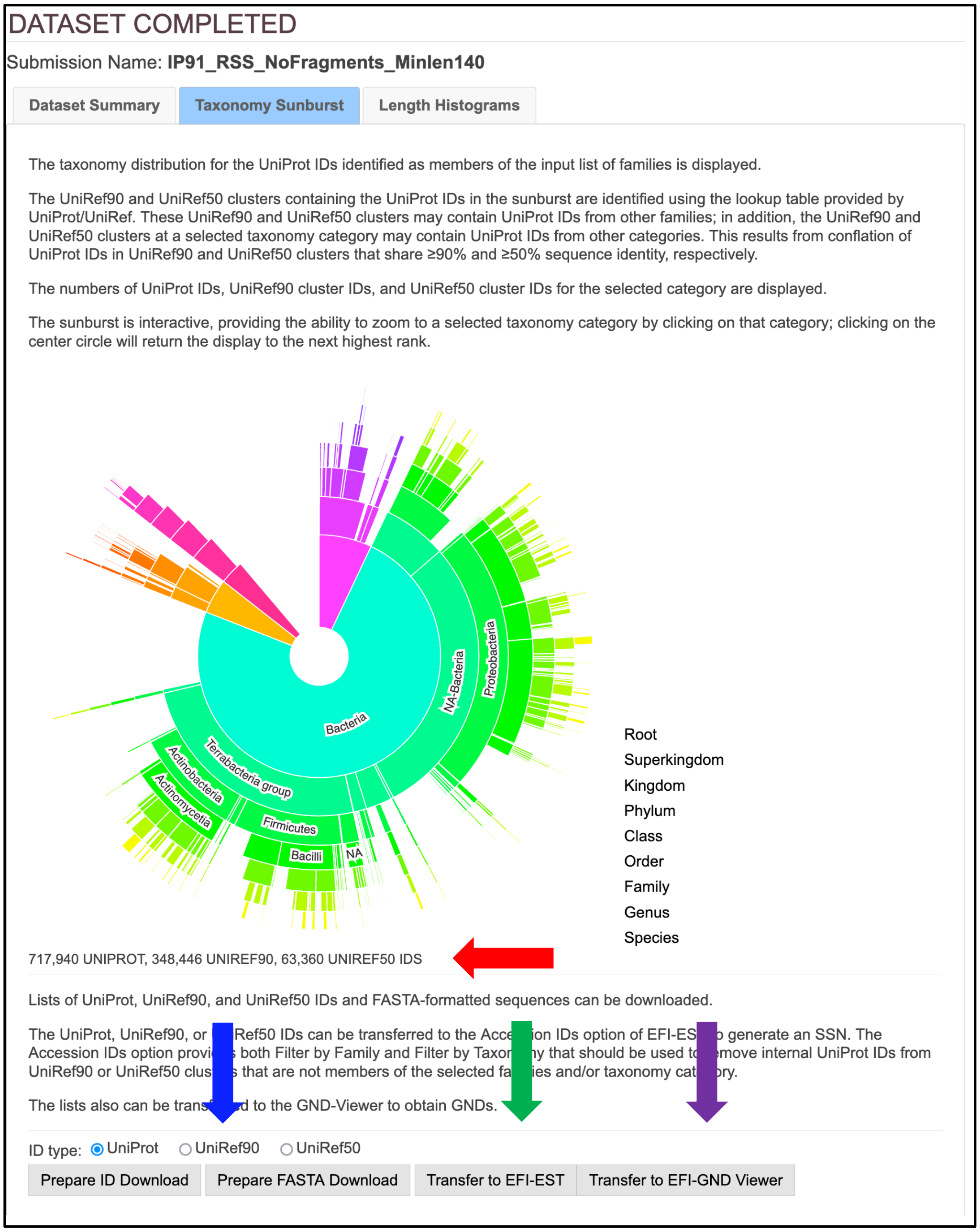
Maximum Length:

Job name: *IP91_RSS_NoFragments_Minlen140* (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_NoFragments_Minlen140**

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 ≥90% and ≥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.

717,940 UNIPROT, 348,446 UNIREF90, 63,360 UNIREF50 IDS

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

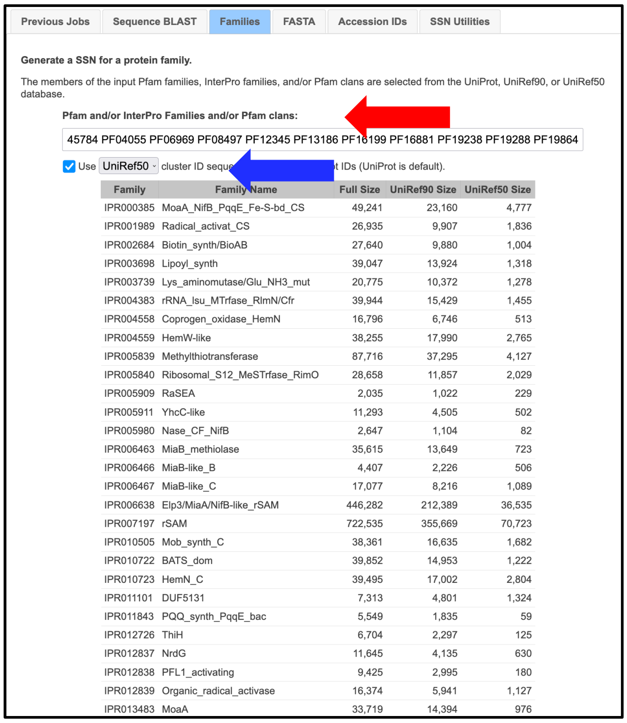
The UniProt, UniRef90, or Ref50 IDs can be transferred to the Acce ion IDs option of EFI-ES o generate an SSN. The Accession IDs option provi s both Filter by Family and Filter by Taxon y that should be used t move internal UniProt IDs from UniRef90 or UniRef50 clus s that are not members of the selected fa ies and/or taxonomy cat ory.

The lists also can be trans d to the GND-Viewer to obtain GNDs.

ID type: ● UniProt ○ UniRef90 ○ UniRef50

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).



Generate a SSN for a protein family.

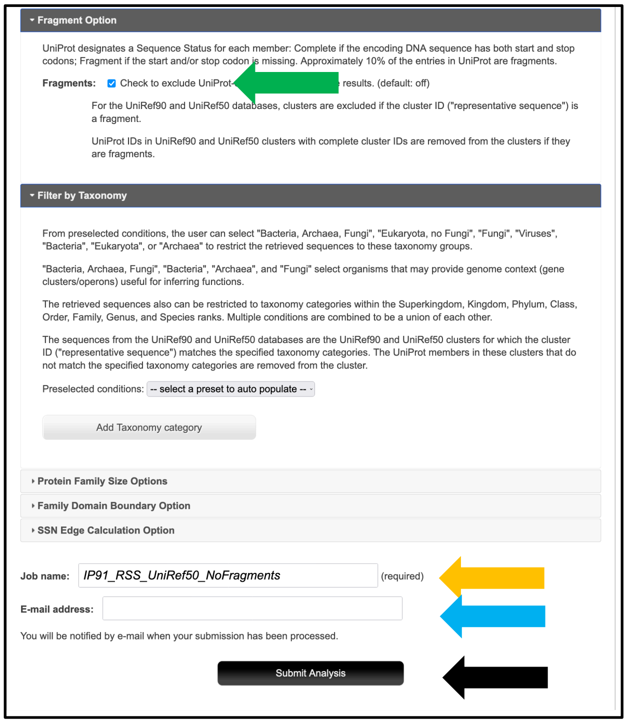
The members of the input Pfam families, InterPro families, and/or Pfam clans are selected from the UniProt, UniRef90, or UniRef50 database.

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

45784 PF04055 PF06969 PF08497 PF12345 PF13186 PF16199 PF16881 PF19238 PF19288 PF19864

☑ Use UniRef50 cluster ID seque t 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_lsu_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
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



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.

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 --

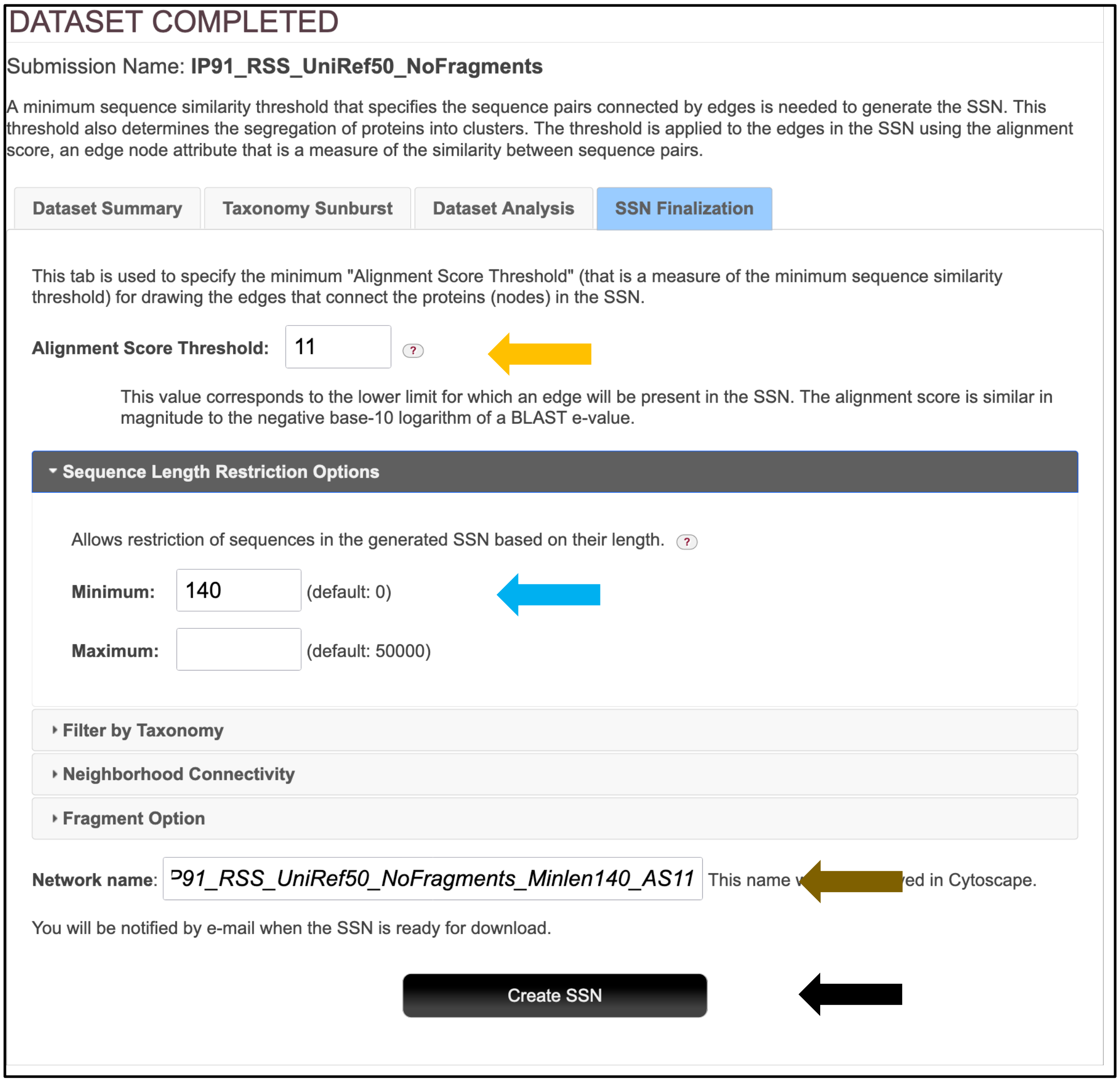
Add Taxonomy category
Protein Family Size Options
Family Domain Boundary Option
SSN Edge Calculation Option

Job name: IP91_RSS_UniRef50_NoFragments (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.

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: 11

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.

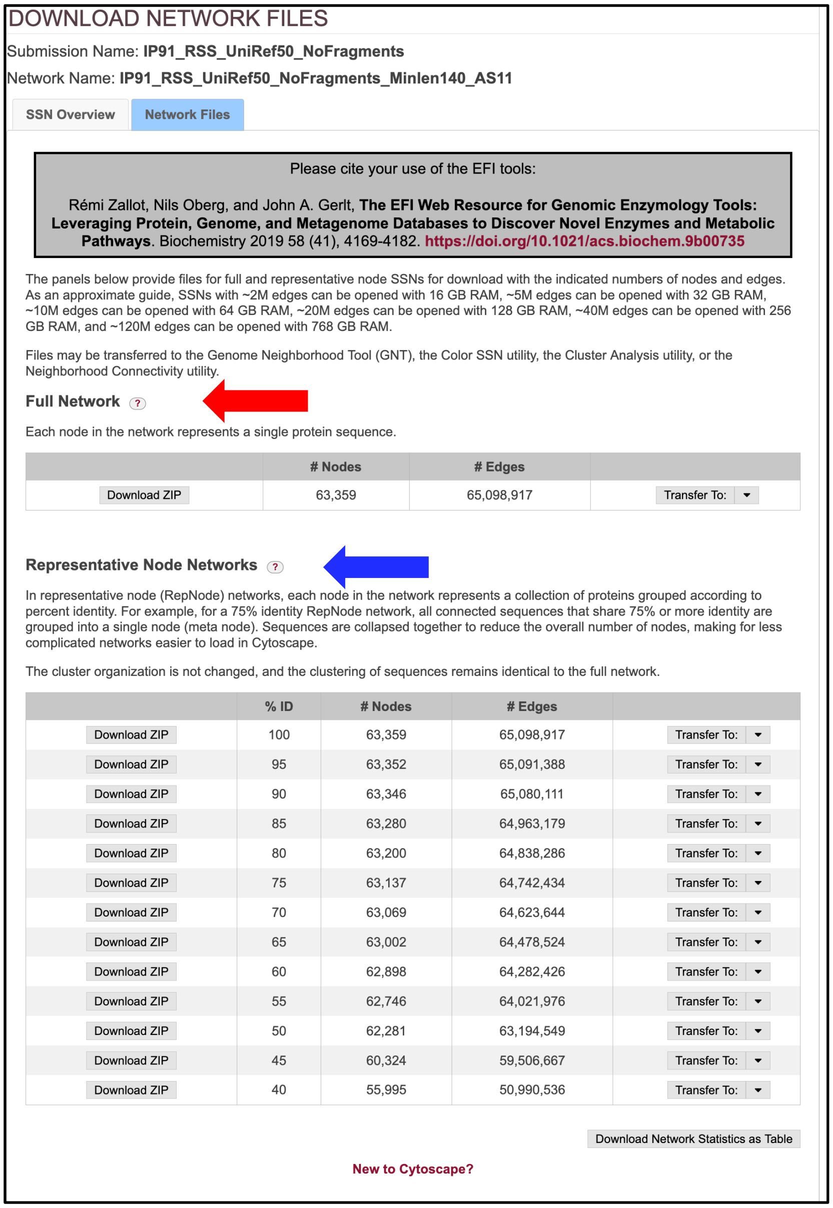
Minimum: 140 (default: 0)

Maximum: (default: 50000)

Network name: P91_RSS_UniRef50_NoFragments_Minlen140_AS11 This name w ed 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 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**

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.

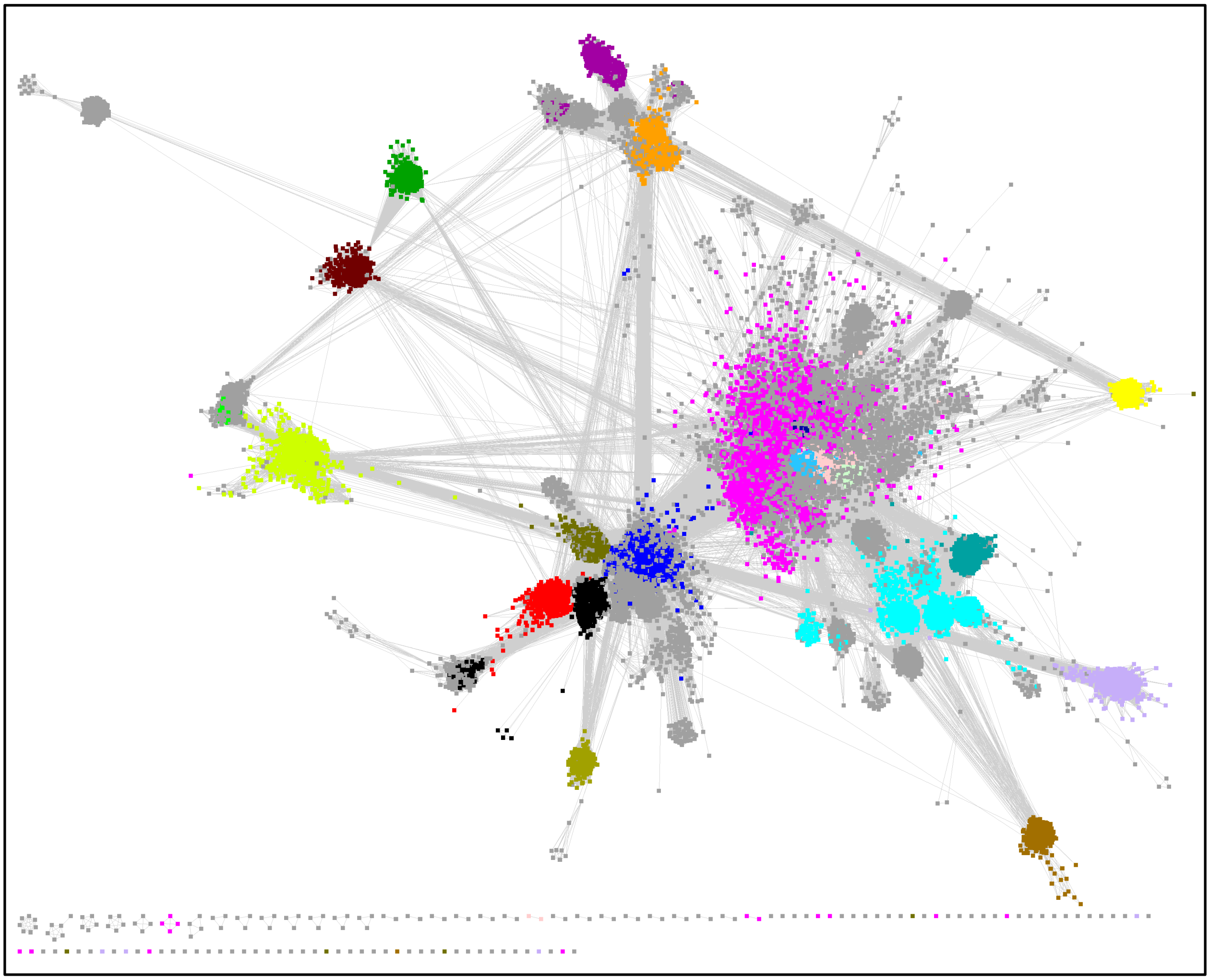
	# Nodes	# Edges	
Download ZIP	63,359	65,098,917	Transfer To: ▾

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.

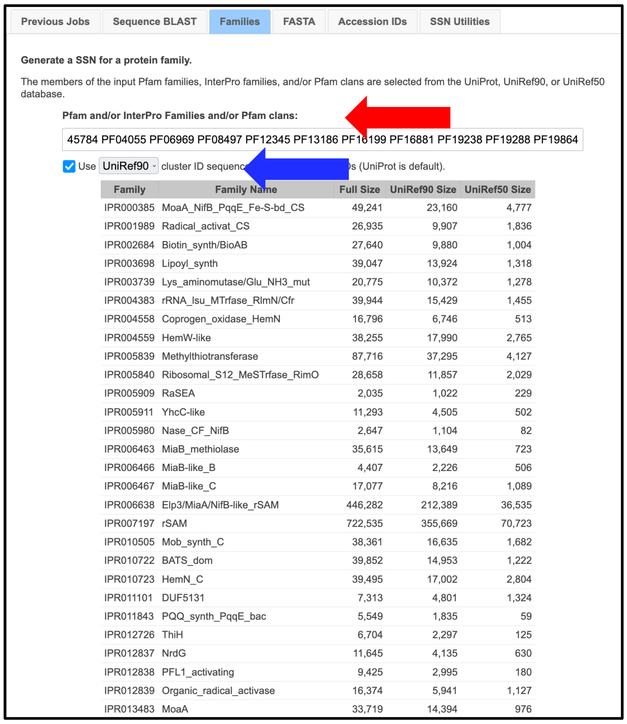
	% ID	# Nodes	# Edges	
Download ZIP	100	63,359	65,098,917	Transfer To: ▾
Download ZIP	95	63,352	65,091,388	Transfer To: ▾
Download ZIP	90	63,346	65,080,111	Transfer To: ▾
Download ZIP	85	63,280	64,963,179	Transfer To: ▾
Download ZIP	80	63,200	64,838,286	Transfer To: ▾
Download ZIP	75	63,137	64,742,434	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: ▾



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).



Generate a SSN for a protein family.

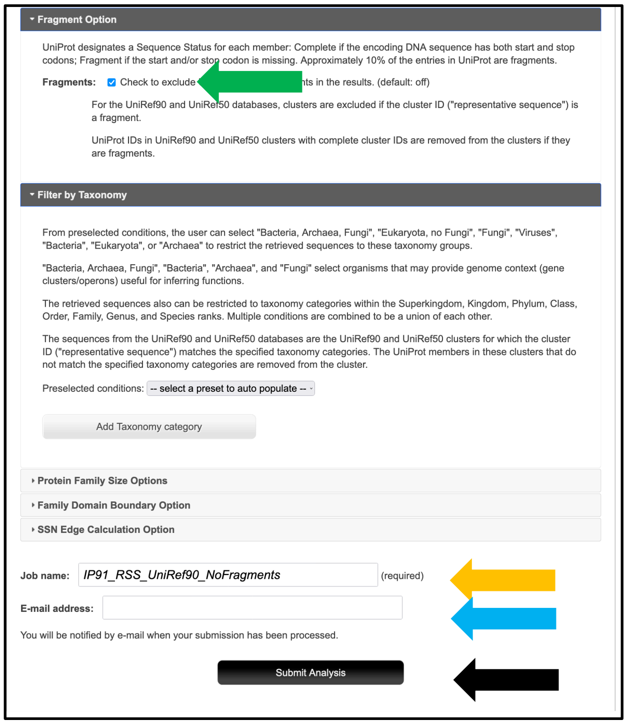
The members of the input Pfam families, InterPro families, and/or Pfam clans are selected from the UniProt, UniRef90, or UniRef50 database.

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

45784 PF04055 PF06969 PF08497 PF12345 PF13186 PF16199 PF16881 PF19238 PF19288 PF19864

☑ Use UniRef90 cluster ID sequenc s (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_lsu_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
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


Fragment Option
Filter by Taxonomy
Add Taxonomy category
Protein Family Size Options
Family Domain Boundary Option
SSN Edge Calculation Option
Submit Analysis

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 nts 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.

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.

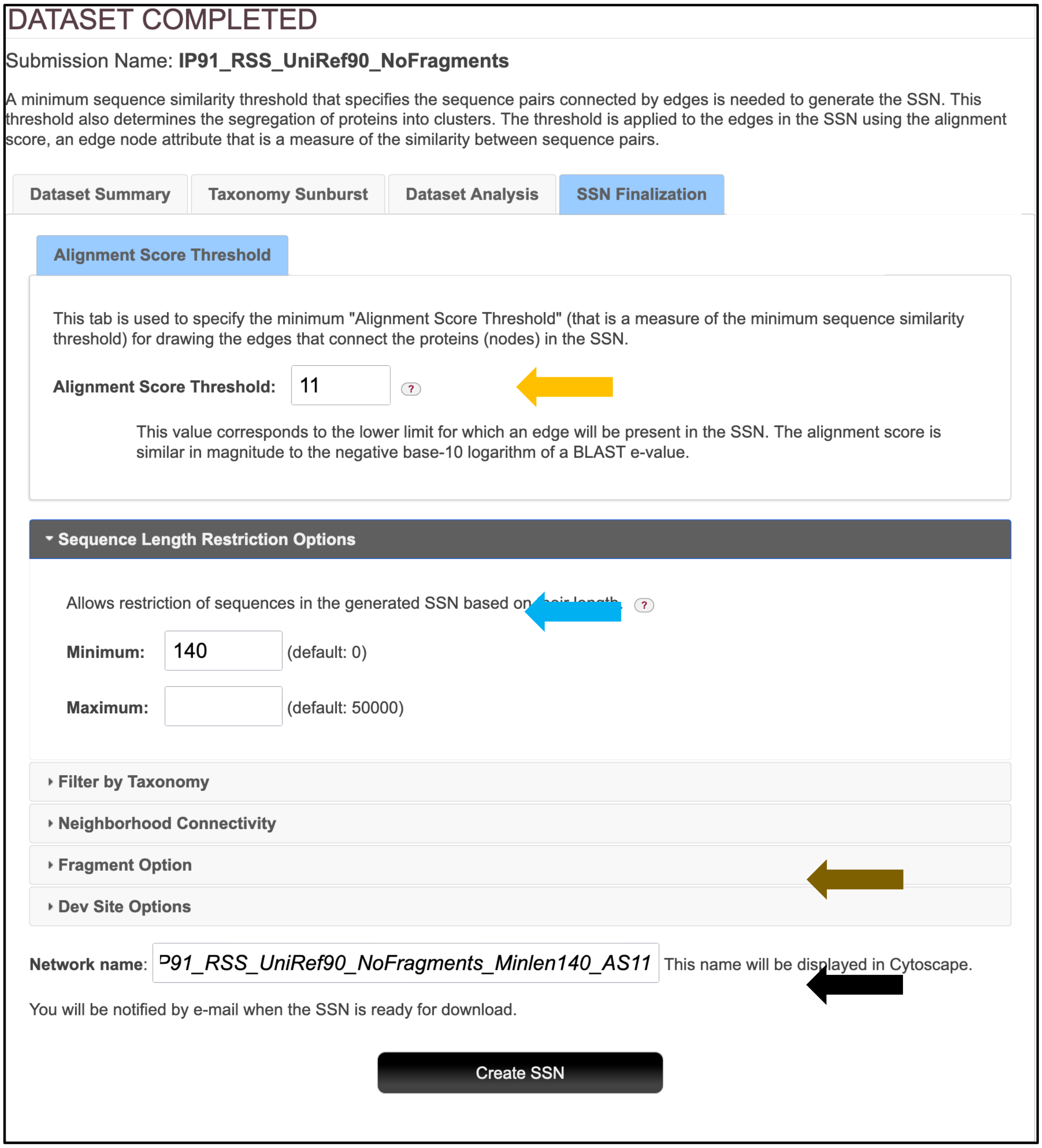
Preselected conditions: -- select a preset to auto populate --

Job name: *IP91_RSS_UniRef90_NoFragments* (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.

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: 11

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.

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

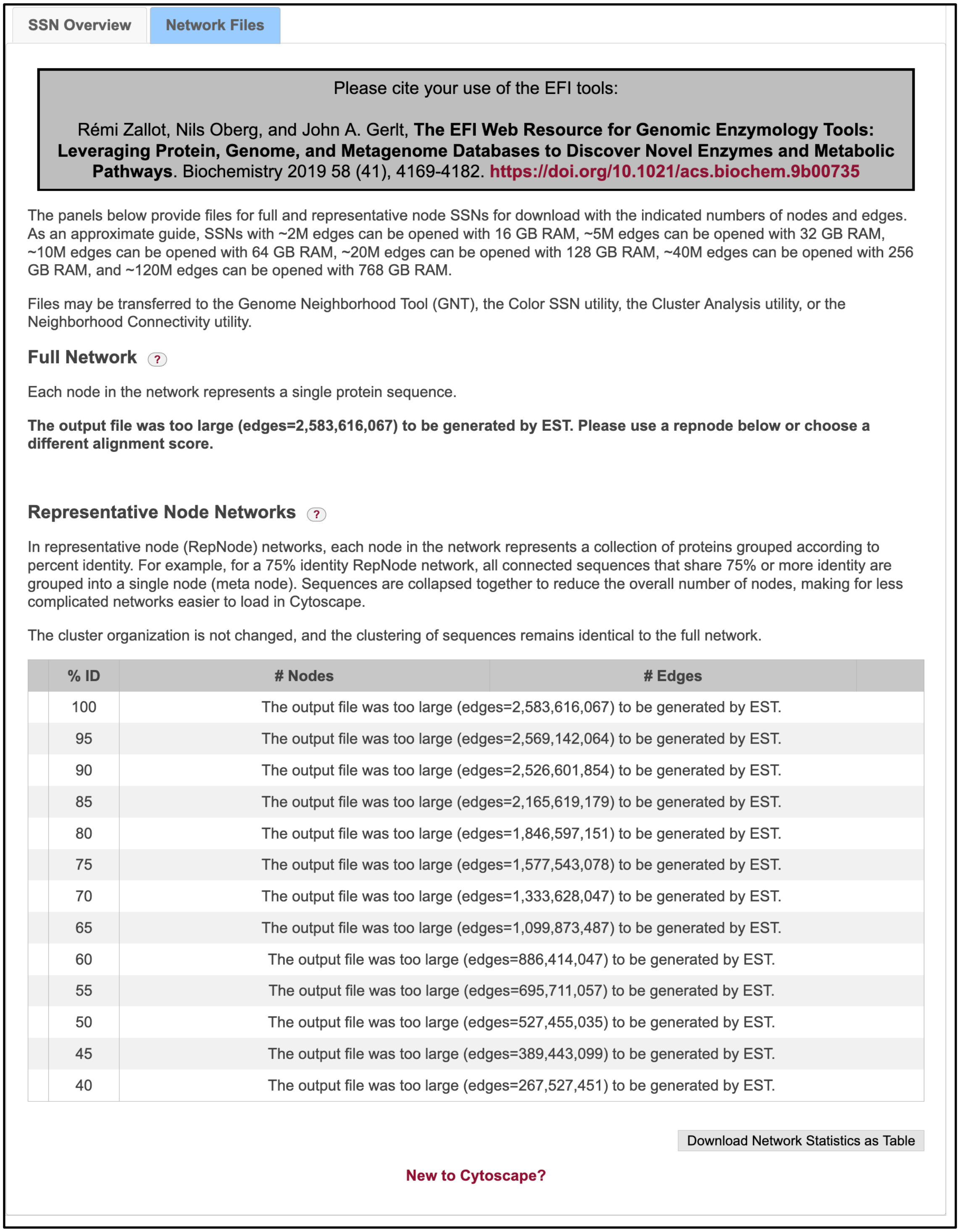
Minimum: 140 (default: 0)

Maximum: (default: 50000)

Network name: P91_RSS_UniRef90_NoFragments_Minlen140_AS11 This name will be displayed in Cytoscape.

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

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.



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 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.

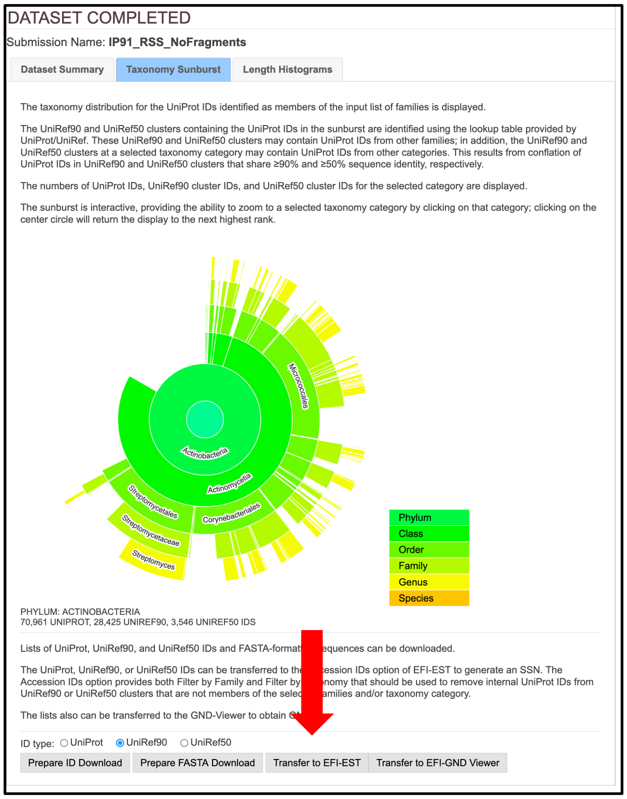
The cluster organization is not changed, and the clustering of sequences remains identical to the full network.

% ID	# Nodes	# Edges
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.	

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).



DATASET COMPLETED

Submission Name: **IP91_RSS_NoFragments**

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 ≥90% and ≥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.

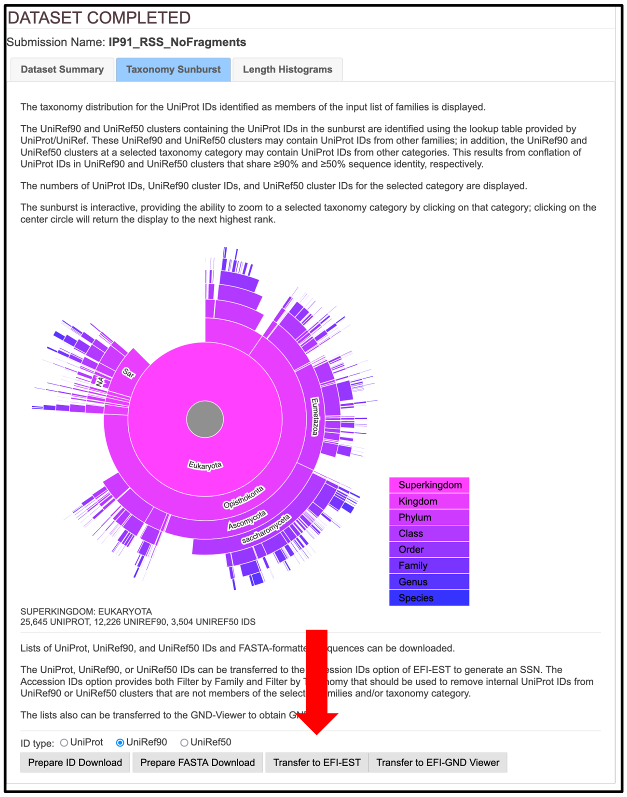
PHYLUM: ACTINOBACTERIA
70,961 UNIPROT, 28,425 UNIREF90, 3,546 UNIREF50 IDS

Lists of UniProt, UniRef90, and UniRef50 IDs and FASTA-format equences can be downloaded.

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

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

ID type: ○ UniProt ● UniRef90 ○ UniRef50



DATASET COMPLETED

Submission Name: **IP91_RSS_NoFragments**

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 ≥90% and ≥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.

SUPERKINGDOM: EUKARYOTA
25,645 UNIPROT, 12,226 UNIREF90, 3,504 UNIREF50 IDS

Lists of UniProt, UniRef90, and UniRef50 IDs and FASTA-formatte quences can be downloaded.

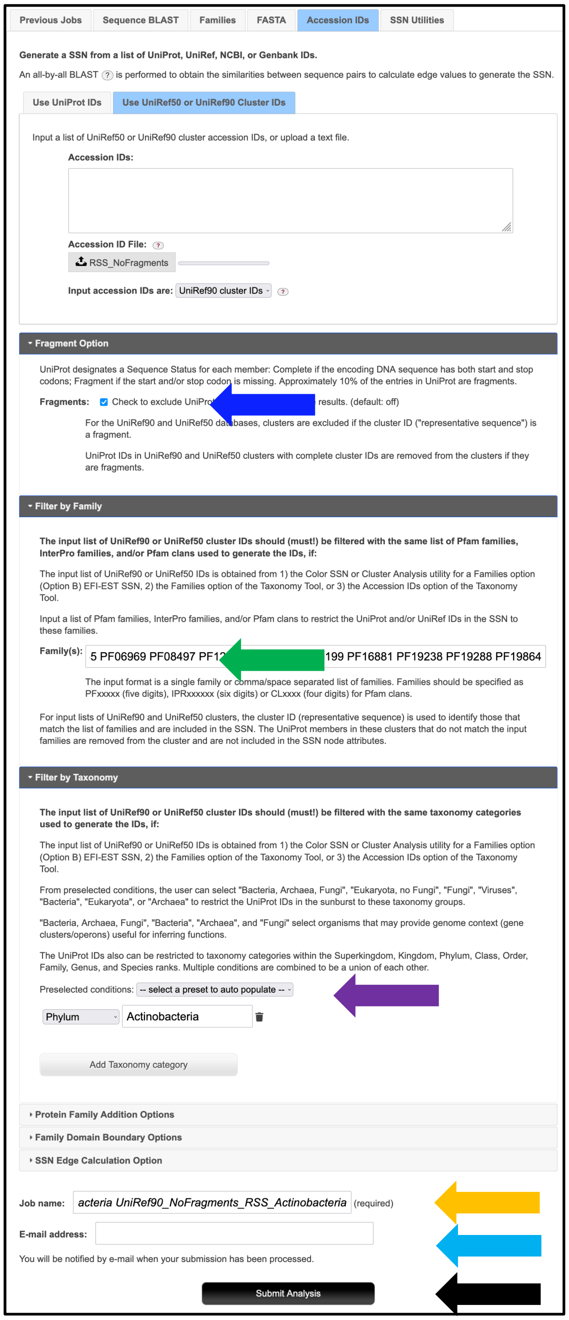
The UniProt, UniRef90, or UniRef50 IDs can be transferred to the ession IDs option of EFI-EST to generate an SSN. The Accession IDs option provides both Filter by Family and Filter by T omy that should be used to remove internal UniProt IDs from UniRef90 or UniRef50 clusters that are not members of the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

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

ID type: ○ UniProt ● UniRef90 ○ UniRef50

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).



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.

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 UniPro 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) EFI-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 PF1 199 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) EFI-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 --

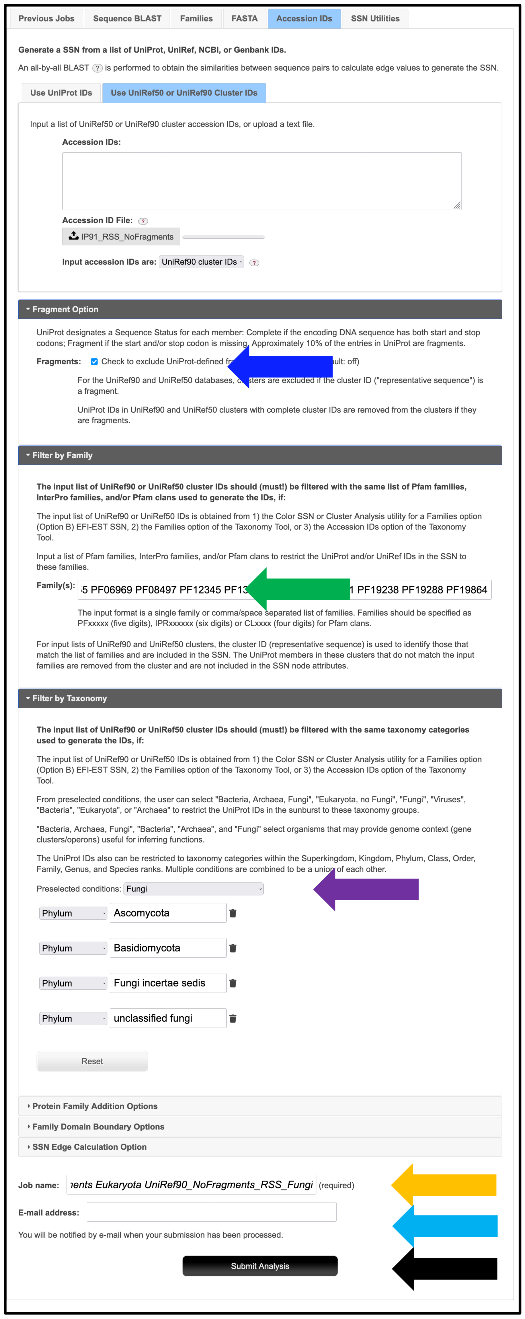
Phylum Actinobacteria

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.



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.

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 fr ult: off)

For the UniRef90 and UniRef50 databases, c sters 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) EFI-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 PF13 1 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) EFI-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

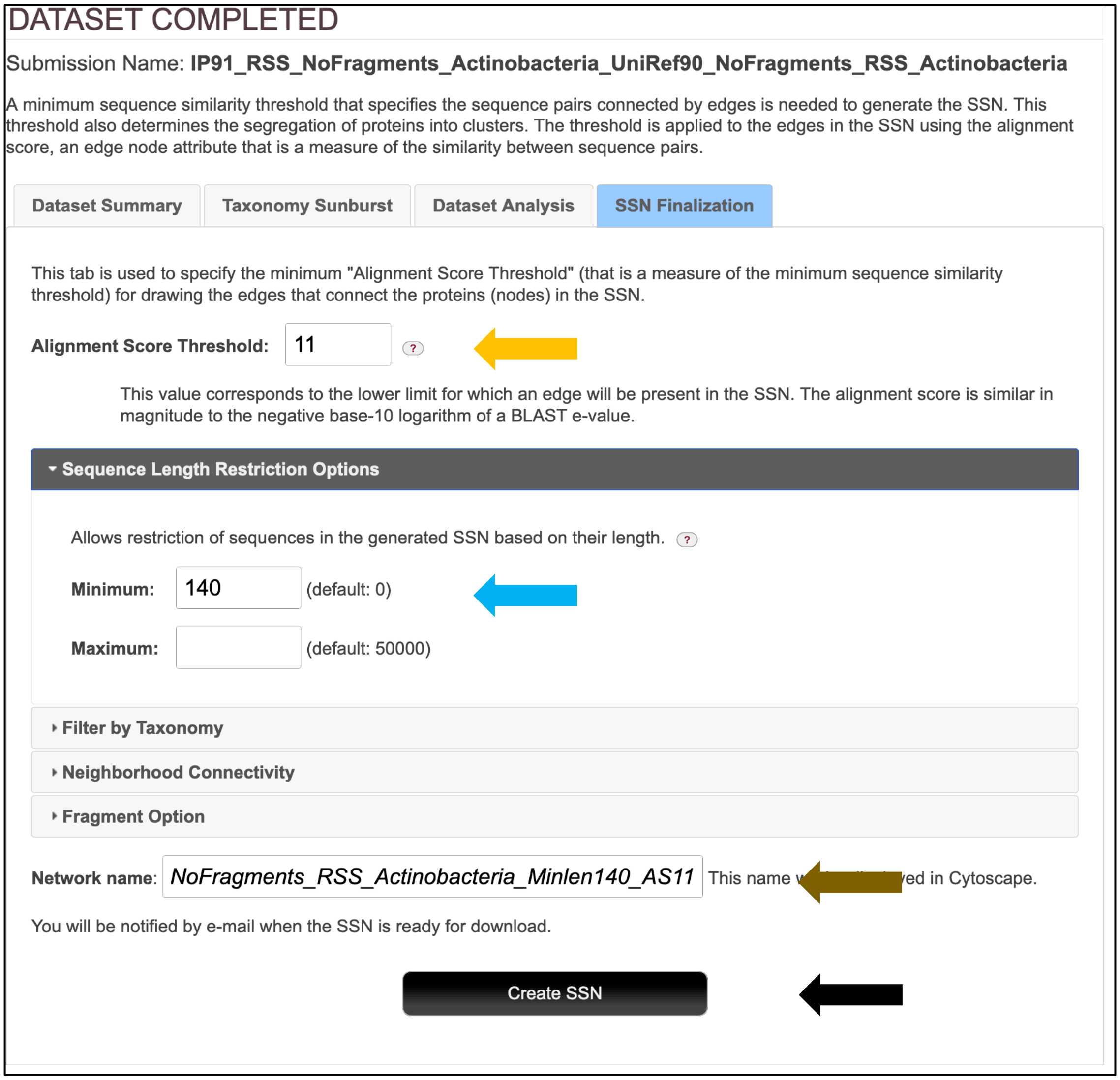
Reset
Protein Family Addition Options
Family Domain Boundary Options
SSN Edge Calculation Option

Job name: *ıents Eukaryota UniRef90_NoFragments_RSS_Fungi* (required)

E-mail address:

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

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.

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: 11

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: 140 (default: 0)

Maximum: (default: 50000)

Filter by Taxonomy

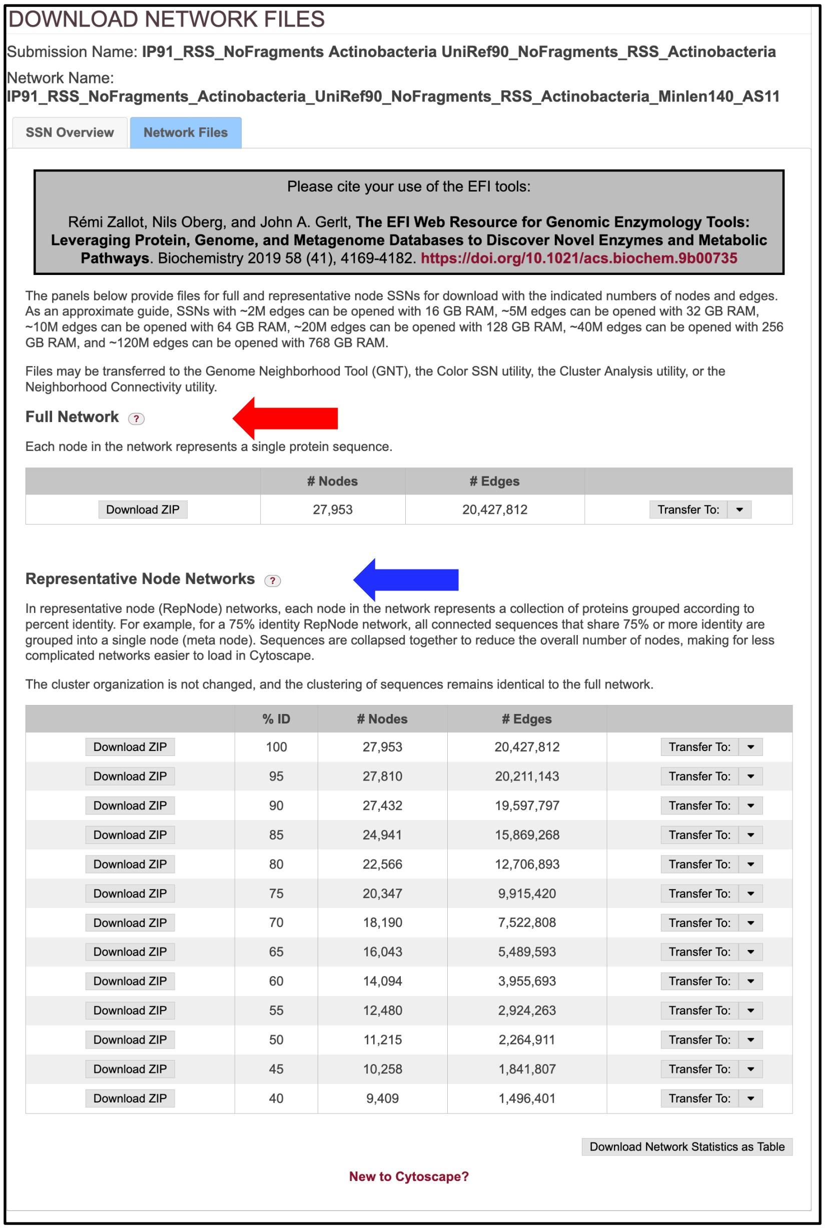
Neighborhood Connectivity

Fragment Option

Network name: *NoFragments_RSS_Actinobacteria_Minlen140_AS11* This name w yed 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 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

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.

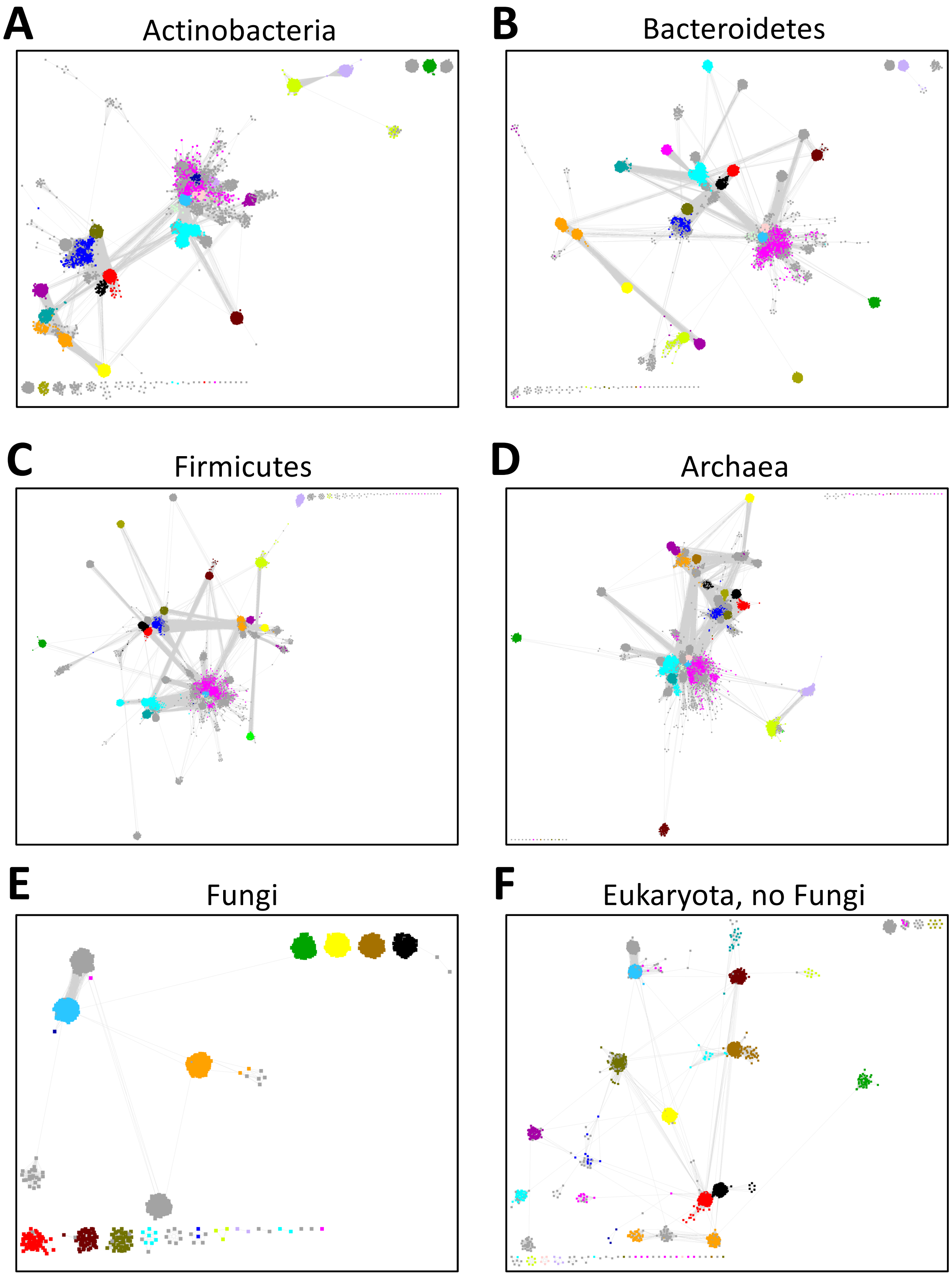
	# Nodes	# Edges	
Download ZIP	27,953	20,427,812	Transfer To: ▾

Representative Node Networks ?

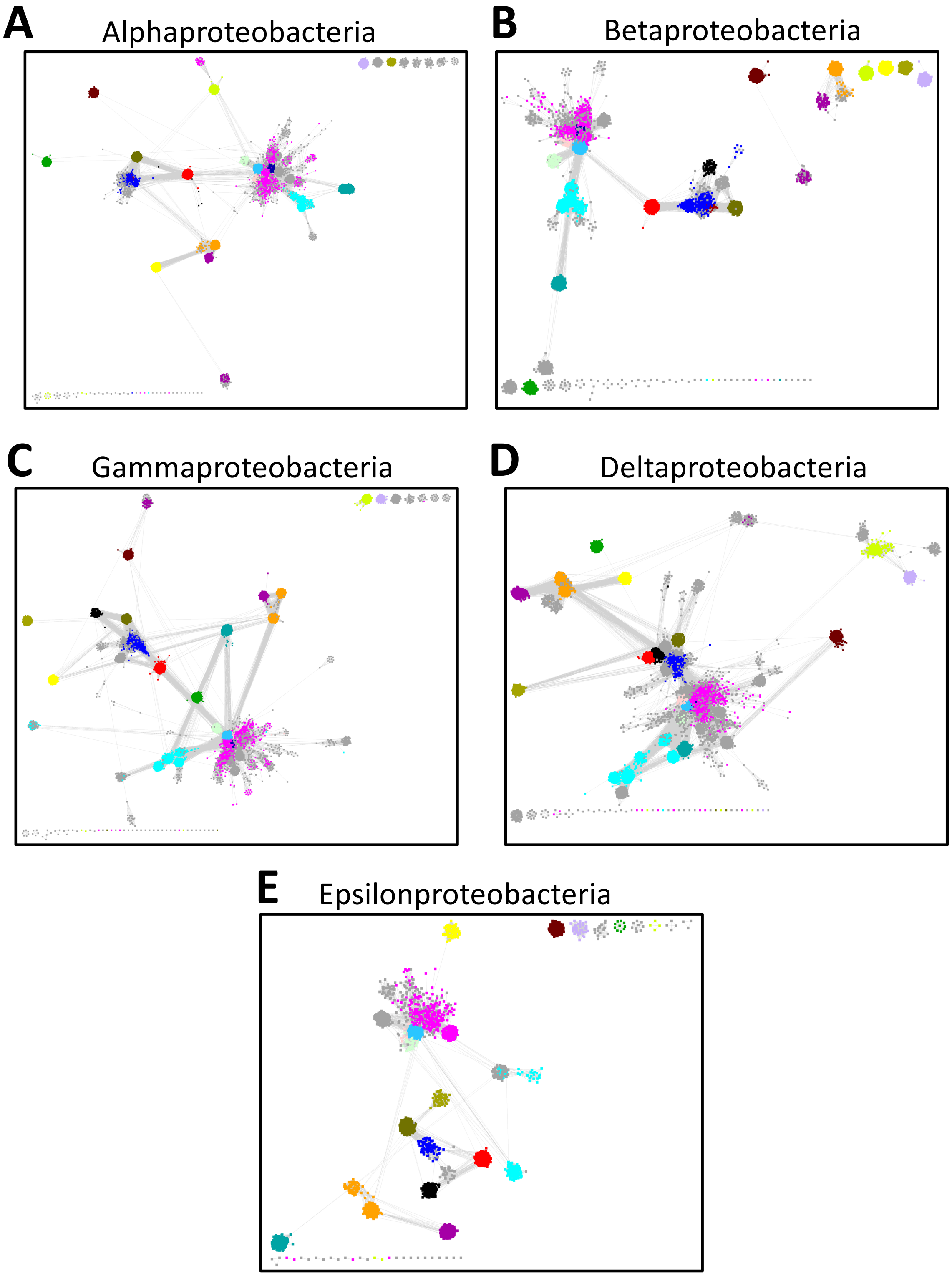
In representative node (RepNode) networks, each node in the network represents a collection of proteins grouped according to percent identity. For example, for a 75% identity RepNode network, all connected sequences that share 75% or more identity are grouped into a single node (meta node). Sequences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


A
Actinobacteria
B
Bacteroidetes
C
Firmicutes
D
Archaea
E
Fungi
F
Eukaryota, no Fungi

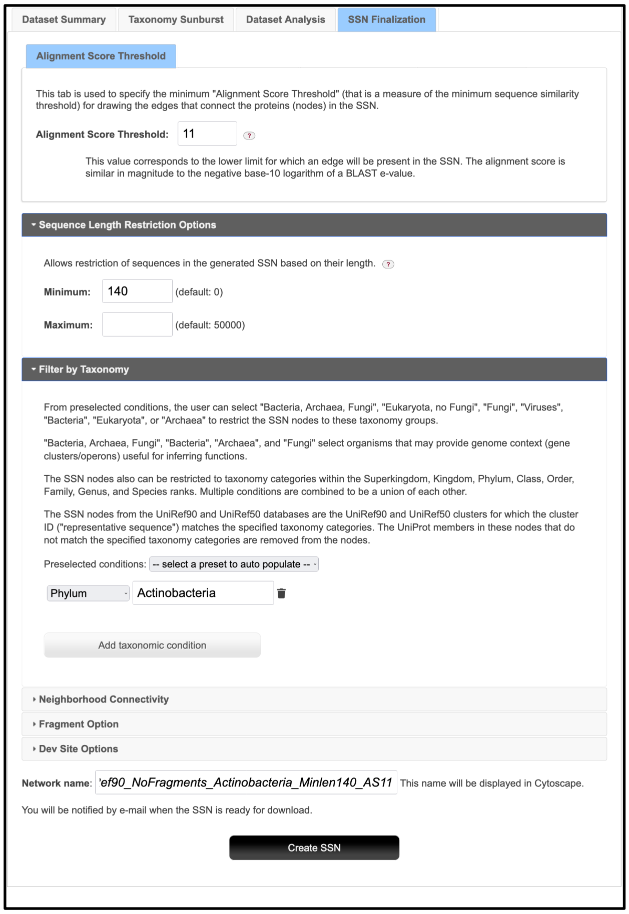
Taxonomy Category-Filtered UniRef 90 Cluster SSNs for the RS Superfamily. The SSNs were generated as described in the text. The SSNs were generated using a minimum length of 140 residues and an alignment score threshold of 11. The nodes are colored according to the subgroups defined by the Structure-Function 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.



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: 11

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: 140 (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.

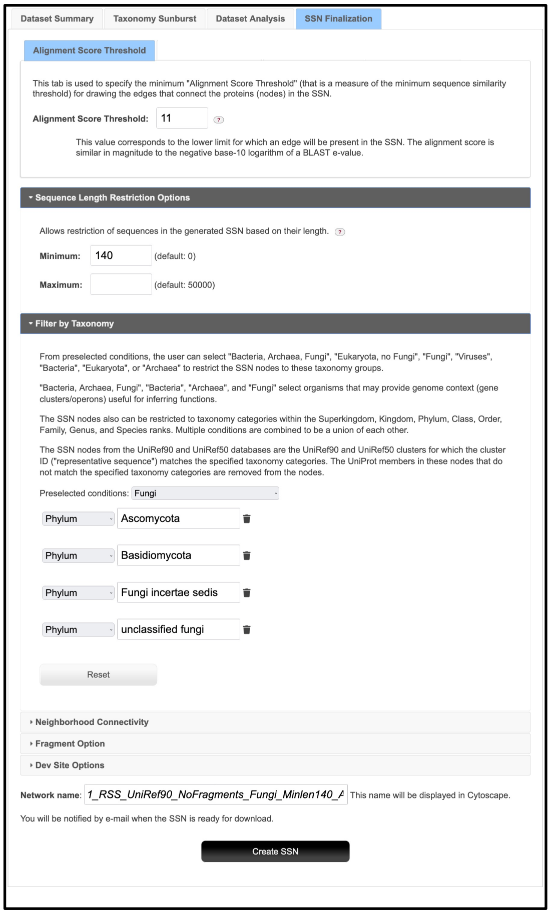
The SSN nodes from the UniRef90 and UniRef50 databases are the UniRef90 and UniRef50 clusters for which the cluster ID ("representative sequence") matches the specified taxonomy categories. The UniProt members in these nodes that do not match the specified taxonomy categories are removed from the nodes.

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

Phylum Actinobacteria

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.



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: 11

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: 140 (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 Ascomycota

Phylum Basidiomycota

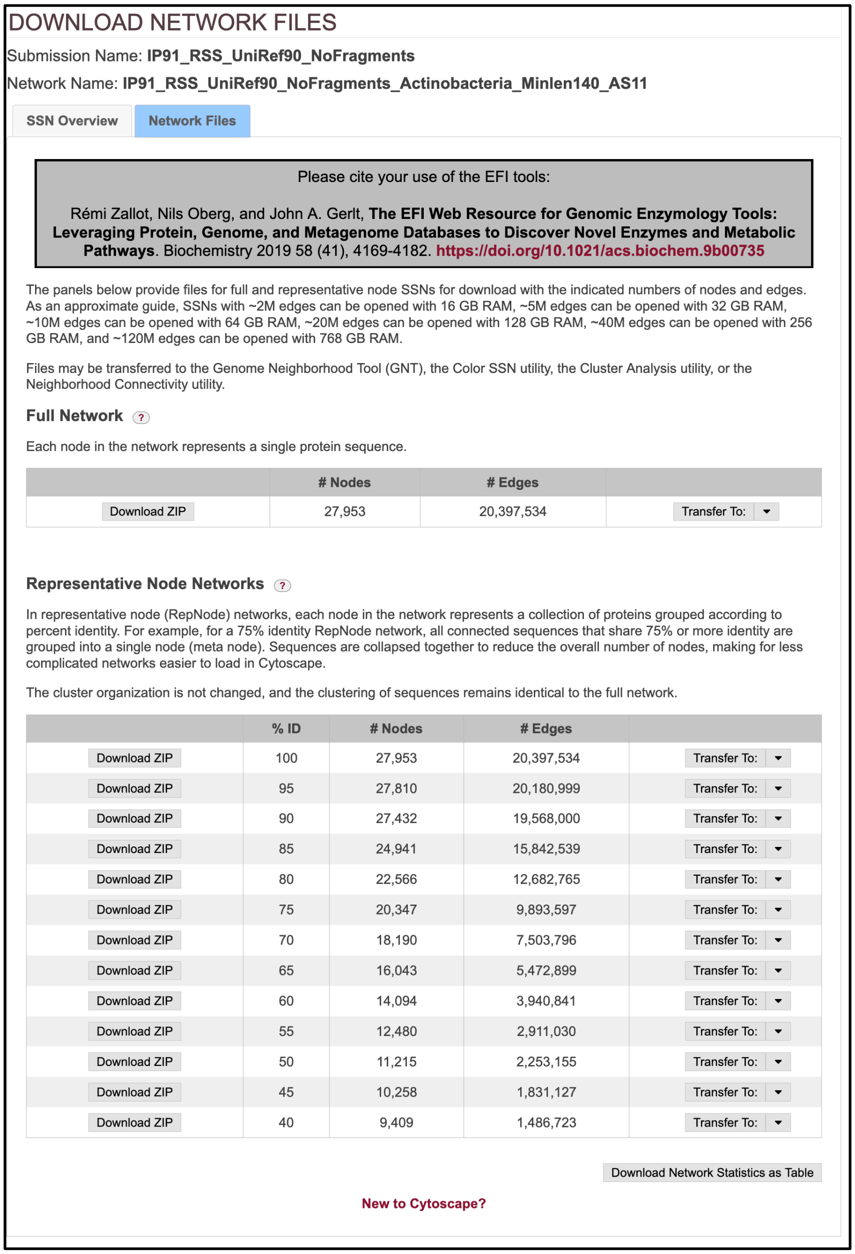
Phylum Fungi incertae sedis

Phylum unclassified fungi

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**

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	27,953	20,397,534	Transfer To: ▾

Representative Node Networks ?

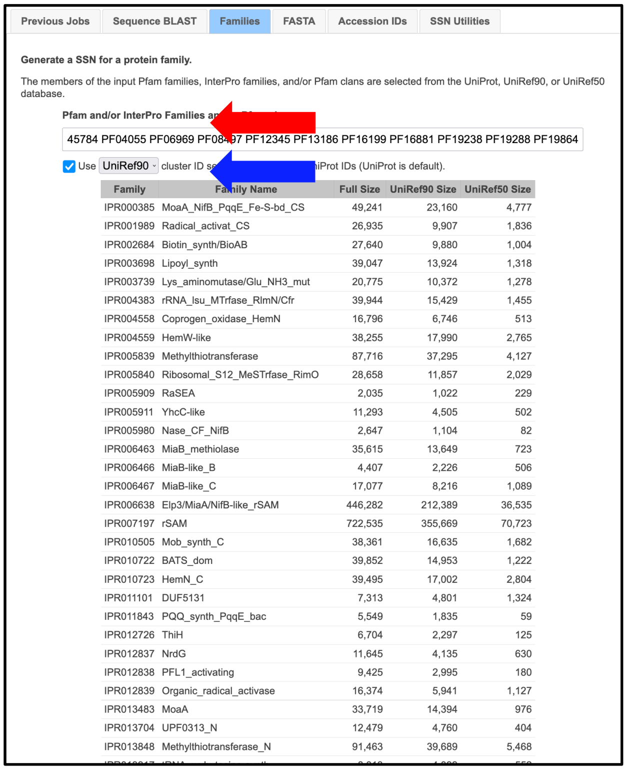
In representative node (RepNode) networks, each node in the network represents a collection of proteins grouped according to percent identity. For example, for a 75% identity RepNode network, all connected sequences that share 75% or more identity are grouped into a single node (meta node). Sequences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

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.

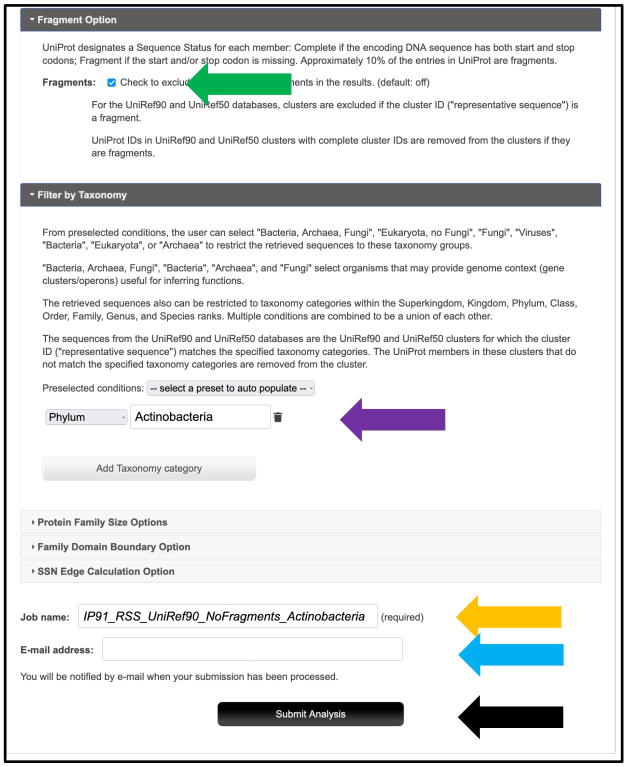
The members of the input Pfam families, InterPro families, and/or Pfam clans are selected from the UniProt, UniRef90, or UniRef50 database.

Pfam and/or InterPro Families an

45784 PF04055 PF06969 PF08497 PF12345 PF13186 PF16199 PF16881 PF19238 PF19288 PF19864

☑ Use UniRef90 cluster ID se iProt 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_lsu_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
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	Methylthiotransferase_N	91,463	39,689	5,468



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 exclud ents 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.

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

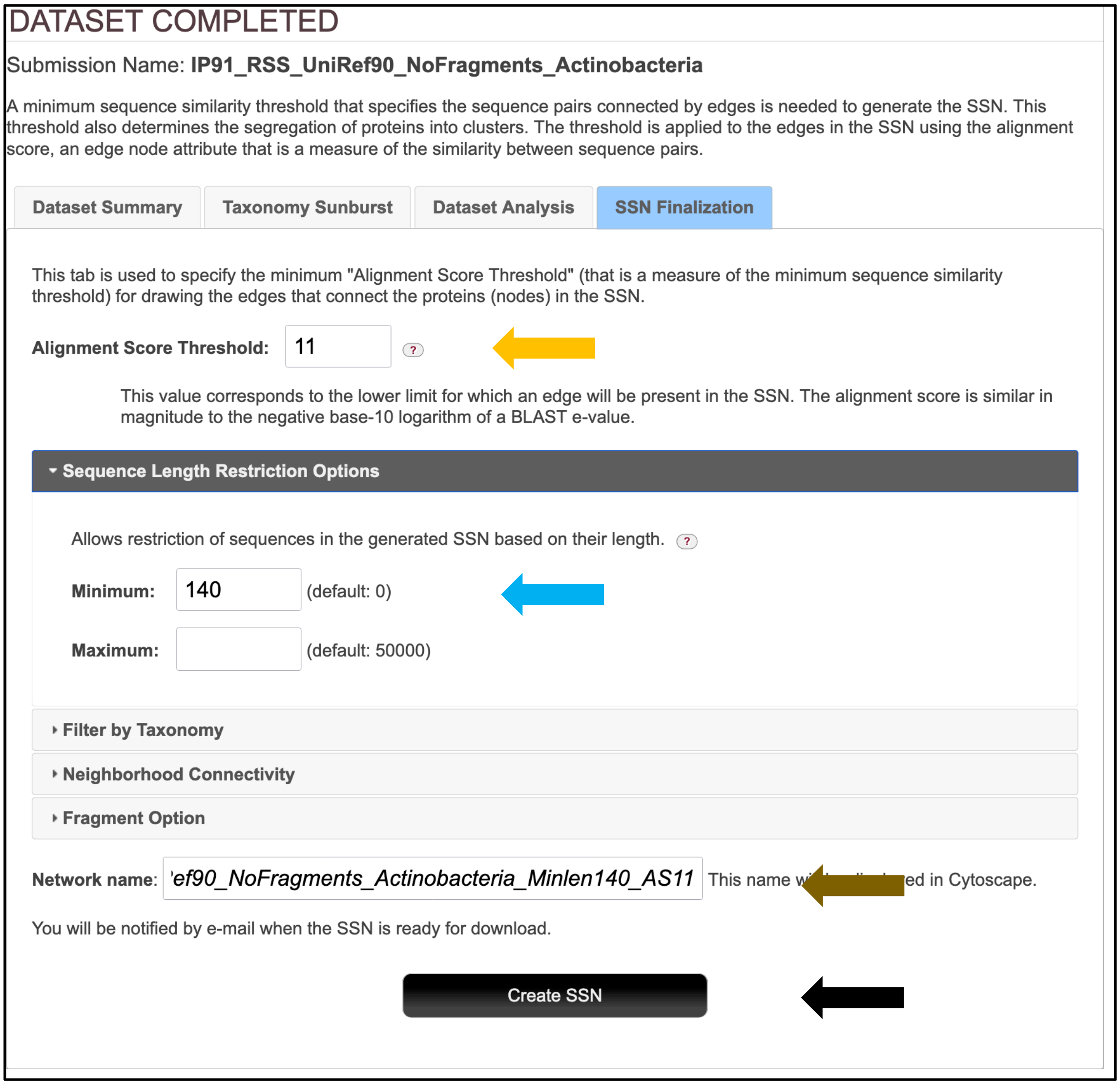
Add Taxonomy category
Protein Family Size Options
Family Domain Boundary Option
SSN Edge Calculation Option

Job name: *IP91_RSS_UniRef90_NoFragments_Actinobacteria* (required)

E-mail address:

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

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.

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: 11

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.

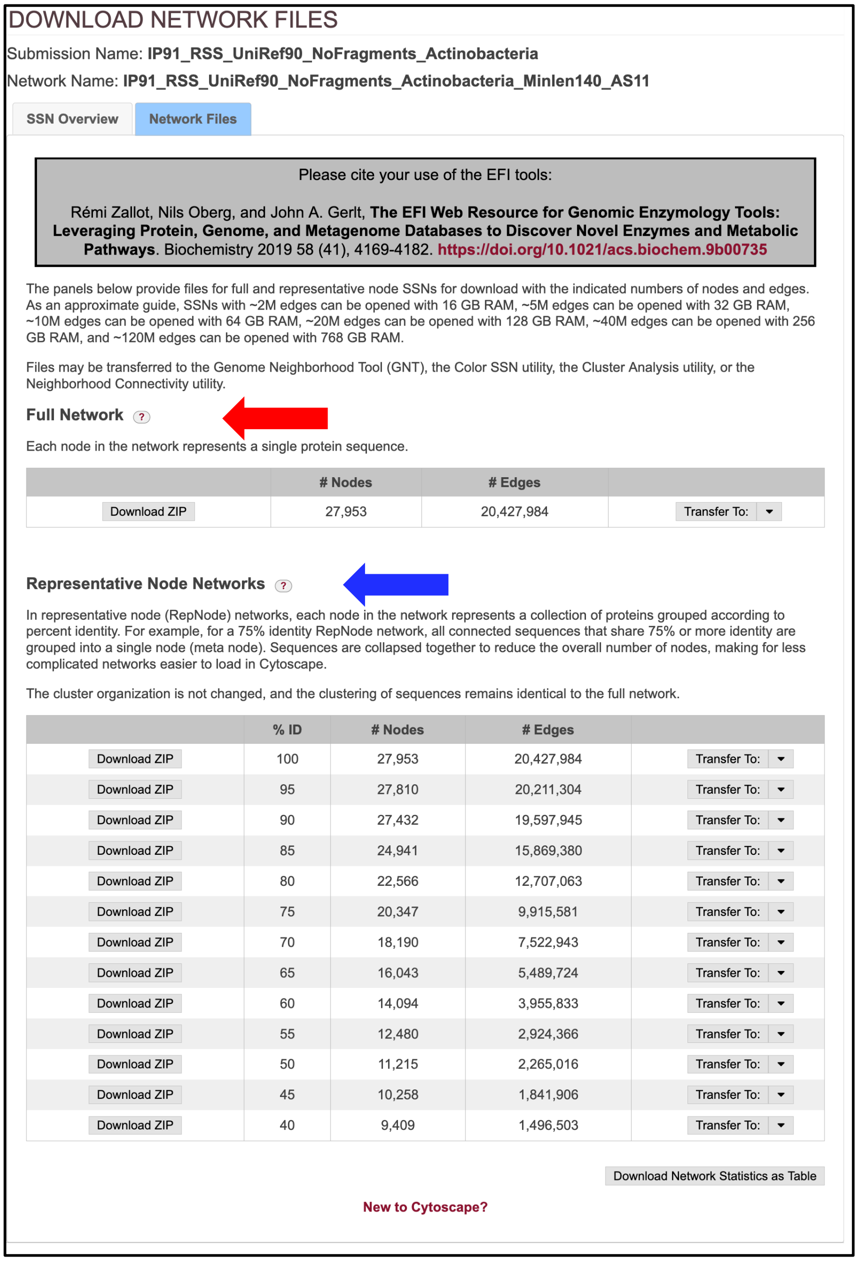
Minimum: 140 (default: 0)

Maximum: (default: 50000)

Network name: 'ef90_NoFragments_Actinobacteria_Minlen140_AS11 This name w[illegible]ed 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 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**

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	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: ▾

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

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

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.