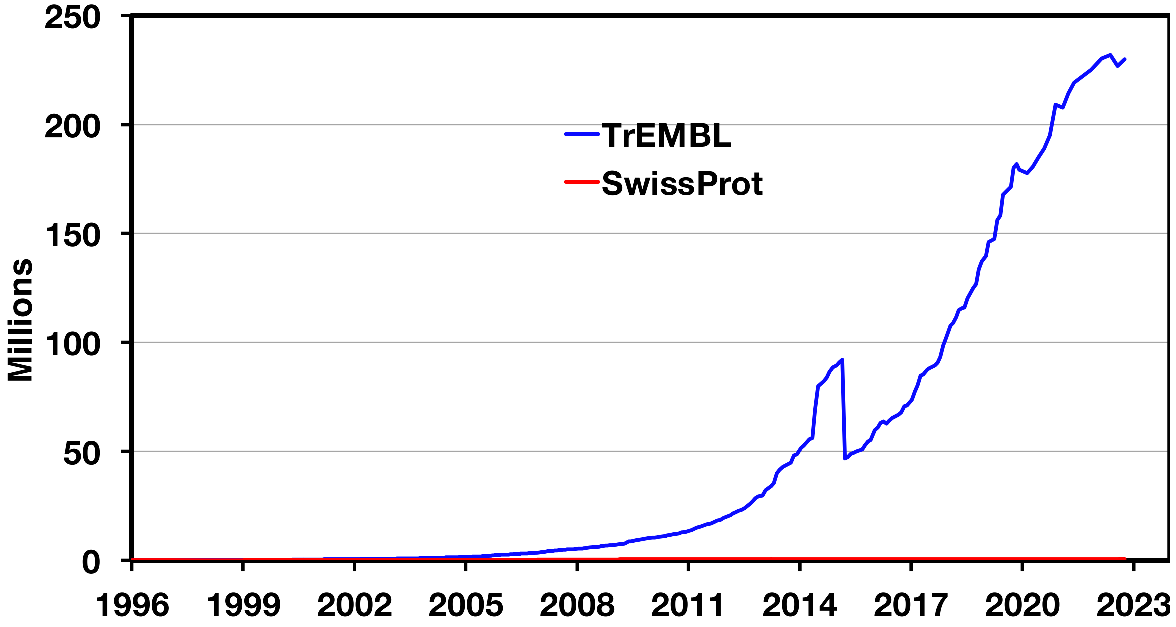
SUPPLEMENTARY FIGURES

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

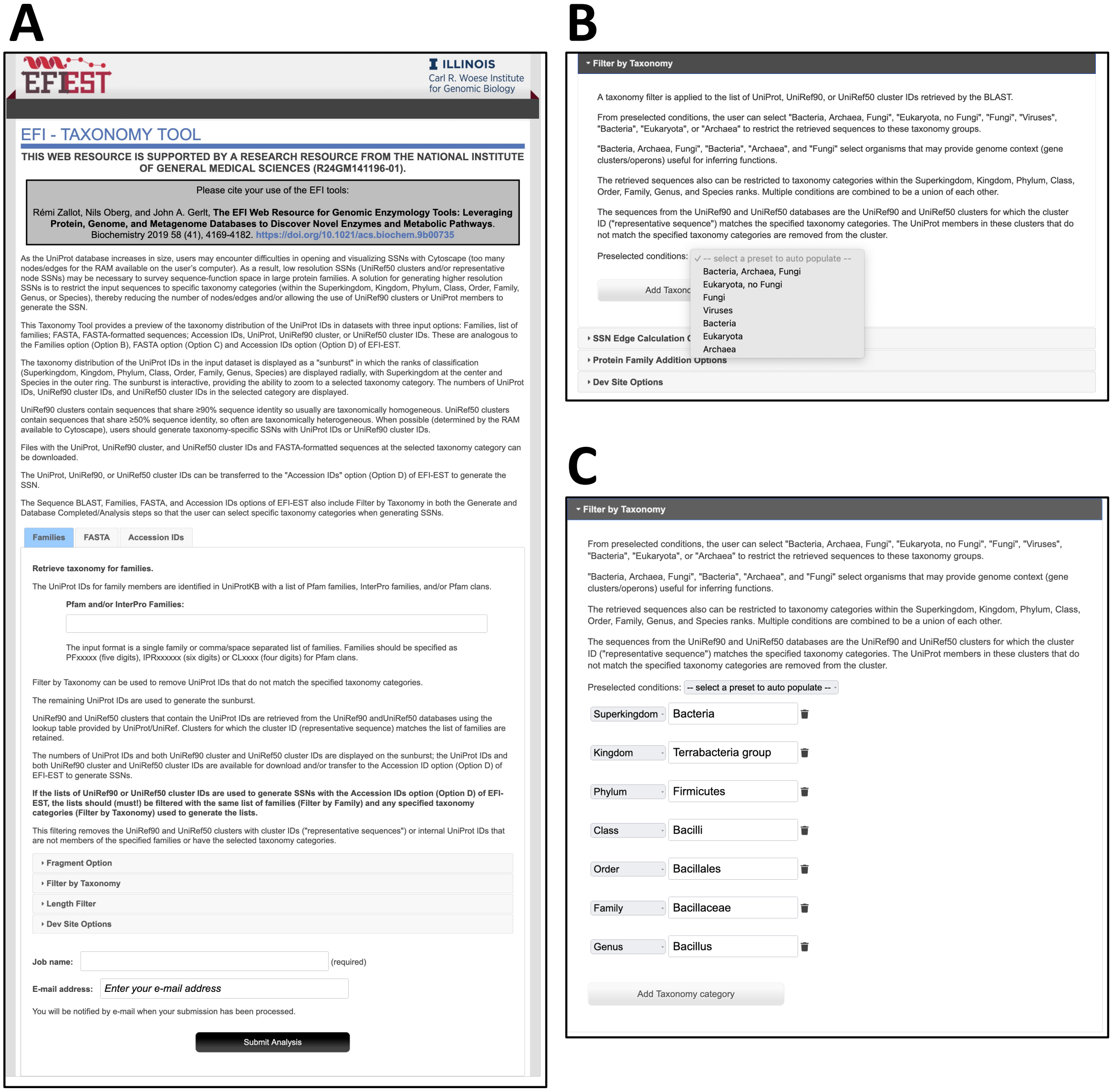
Nils Oberg[1], Rémi Zallot[2,3], and John A. Gerlt[1,4,5*]

[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



Supplementary Figure S1. **Growth of the UniProtKB database**. UniProtKB is the aggregate of the UniProtKB/TrEMBL database that contains computationally annotated entries [229,928,140 in Release 2022_04 (12-October-2022)] and the UniProtKB/SwissProt database that contains manually curated entries [(568,363 in Release 2022_04 (12-October-2022)]. The decrease in the number of entries in 2015 and the more modest increase in recent months is the explained by the inclusion of entries from reference proteomes to manage the growth of the database.



EFI - TAXONOMY TOOL

THIS WEB RESOURCE IS SUPPORTED BY A RESEARCH RESOURCE FROM THE NATIONAL INSTITUTE OF GENERAL MEDICAL SCIENCES (R24GM141196-01).

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

As the UniProt database increases in size, users may encounter difficulties in opening and visualizing SSNs with Cytoscape (too many nodes/edges for the RAM available on the user's computer). As a result, low resolution SSNs (UniRef50 clusters and/or representative node SSNs) may be necessary to survey sequence-function space in large protein families. A solution for generating higher resolution SSNs is to restrict the input sequences to specific taxonomy categories (within the Superkingdom, Kingdom, Phylum, Class, Order, Family, Genus, or Species), thereby reducing the number of nodes/edges and/or allowing the use of UniRef90 clusters or UniProt members to generate the SSN.

This Taxonomy Tool provides a preview of the taxonomy distribution of the UniProt IDs in datasets with three input options: Families, list of families; FASTA, FASTA-formatted sequences; Accession IDs, UniProt, UniRef90 cluster, or UniRef50 cluster IDs. These are analogous to the Families option (Option B), FASTA option (Option C) and Accession IDs option (Option D) of EFI-EST.

The taxonomy distribution of the UniProt IDs in the input dataset is displayed as a "sunburst" in which the ranks of classification (Superkingdom, Kingdom, Phylum, Class, Order, Family, Genus, Species) are displayed radially, with Superkingdom at the center and Species in the outer ring. The sunburst is interactive, providing the ability to zoom to a selected taxonomy category. The numbers of UniProt IDs, UniRef90 cluster IDs, and UniRef50 cluster IDs in the selected category are displayed.

UniRef90 clusters contain sequences that share ≥90% sequence identity so usually are taxonomically homogeneous. UniRef50 clusters contain sequences that share ≥50% sequence identity, so often are taxonomically heterogeneous. When possible (determined by the RAM available to Cytoscape), users should generate taxonomy-specific SSNs with UniProt IDs or UniRef90 cluster IDs.

Files with the UniProt, UniRef90 cluster, and UniRef50 cluster IDs and FASTA-formatted sequences at the selected taxonomy category can be downloaded.

The UniProt, UniRef90, or UniRef50 cluster IDs can be transferred to the "Accession IDs" option (Option D) of EFI-EST to generate the SSN.

The Sequence BLAST, Families, FASTA, and Accession IDs options of EFI-EST also include Filter by Taxonomy in both the Generate and Database Completed/Analysis steps so that the user can select specific taxonomy categories when generating SSNs.

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:

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.

Job name: (required)

E-mail address: *Enter your e-mail address*

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

A taxonomy filter is applied to the list of UniProt, UniRef90, or UniRef50 cluster IDs retrieved by the BLAST.

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:

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.

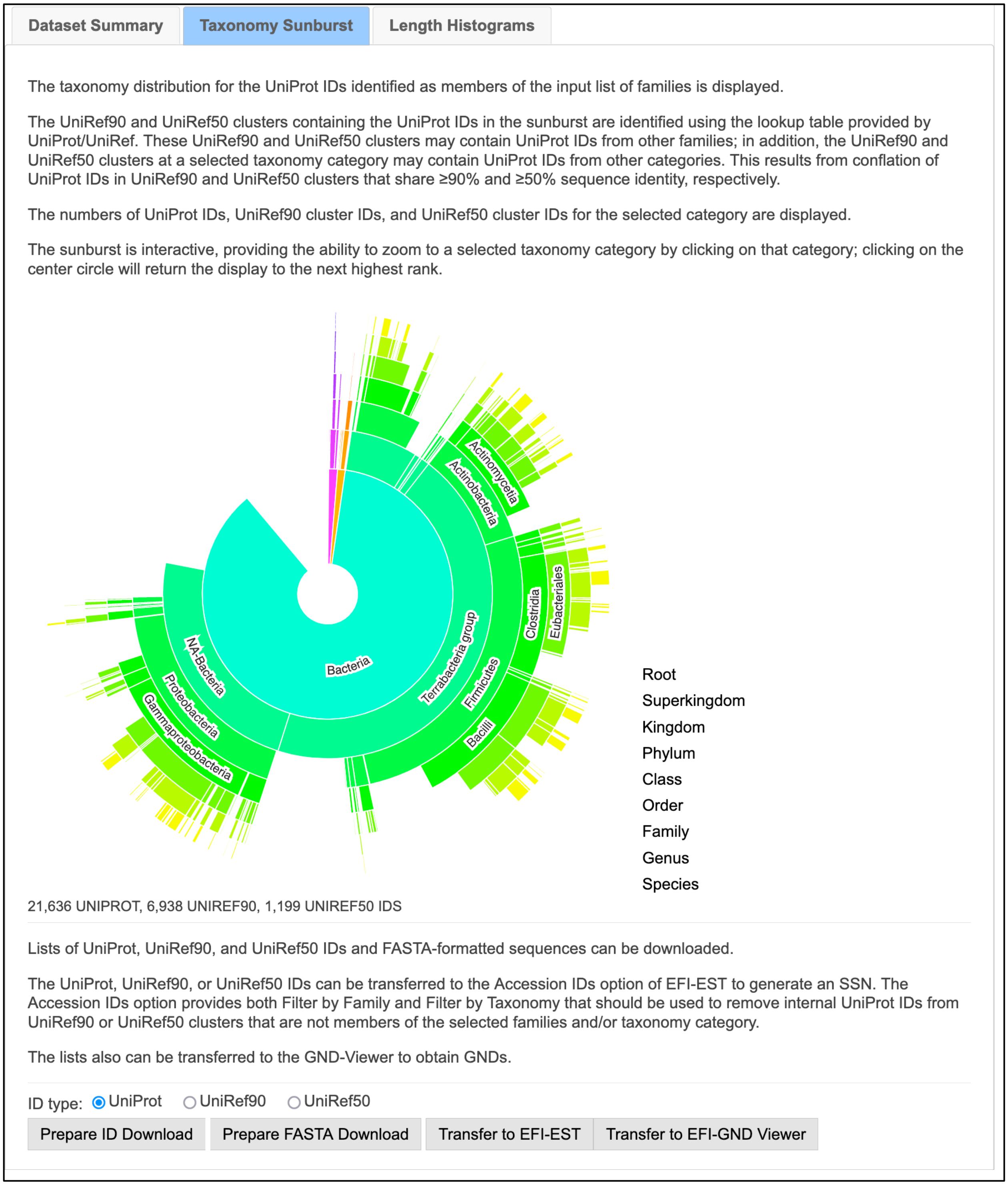
The retrieved sequences also can be restricted to taxonomy categories within the Superkingdom, Kingdom, Phylum, Class, Order, Family, Genus, and Species ranks. Multiple conditions are combined to be a union of each other.

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 --

Superkingdom	Bacteria
Kingdom	Terrabacteria group
Phylum	Firmicutes
Class	Bacilli
Order	Bacillales
Family	Bacillaceae
Genus	Bacillus

Supplementary Figure S2. Panel A, Taxonomy Tool. Panel B, Filter by Taxonomy option in the **Taxonomy Tool**, showing **Preselected conditions**. **Panel C, Filter by Taxonomy** option in the **Taxonomy Tool**, showing **Add Taxonomy Category**.



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.

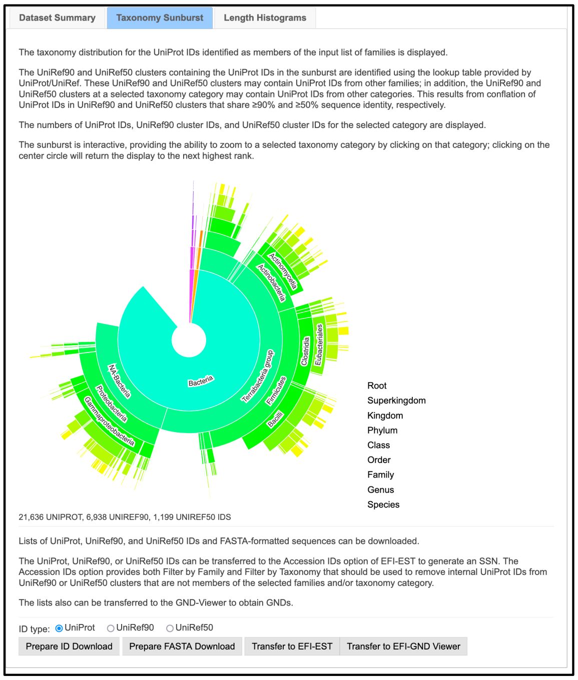
The UniProt, UniRef90, or UniRef50 IDs can be transferred to the Accession IDs option of EFI-EST to generate an SSN. The Accession IDs option provides both Filter by Family and Filter by Taxonomy that should be used to remove internal UniProt IDs from UniRef90 or UniRef50 clusters that are not members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

Supplementary Figure S3. Taxonomy Sunburst.

A All Categories



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

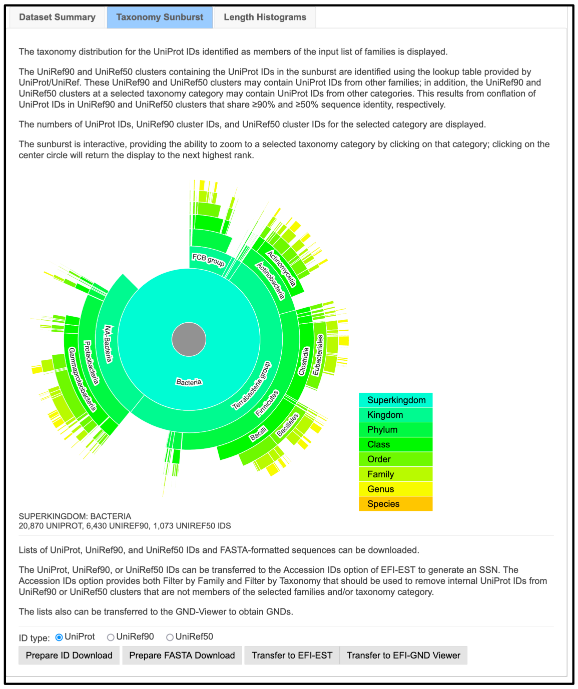
Lists of UniProt, UniRef90, and UniRef50 IDs and FASTA-formatted sequences can be downloaded.

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

B Superkingdom Bacteria



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: BACTERIA
20,870 UNIPROT, 6,430 UNIREF90, 1,073 UNIREF50 IDS

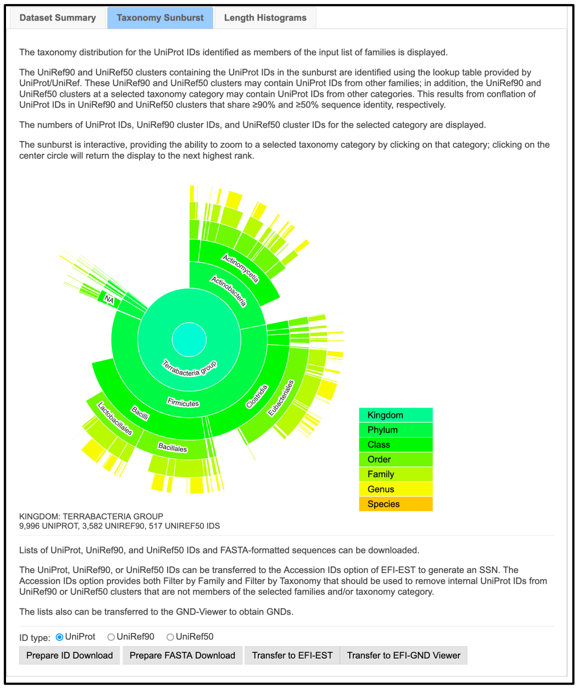
Lists of UniProt, UniRef90, and UniRef50 IDs and FASTA-formatted sequences can be downloaded.

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

C Kingdom Terrabacteria Group



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.

KINGDOM: TERRABACTERIA GROUP
9,996 UNIPROT, 3,582 UNIREF90, 517 UNIREF50 IDS

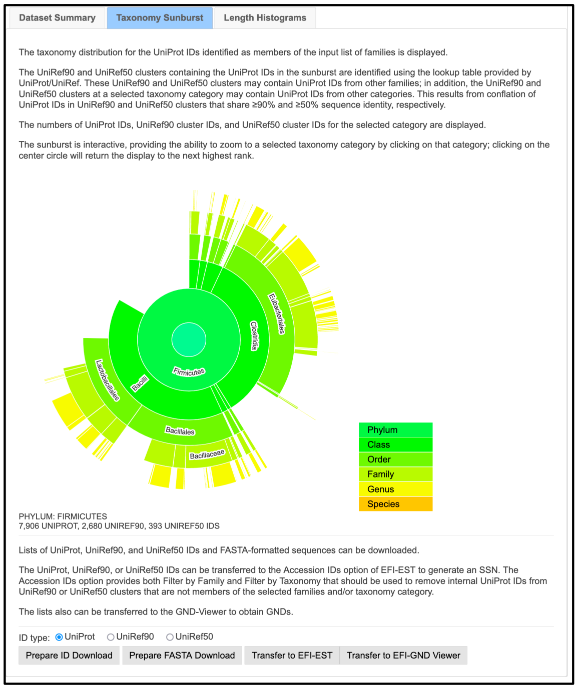
Lists of UniProt, UniRef90, and UniRef50 IDs and FASTA-formatted sequences can be downloaded.

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

D Phylum Firmicutes



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.

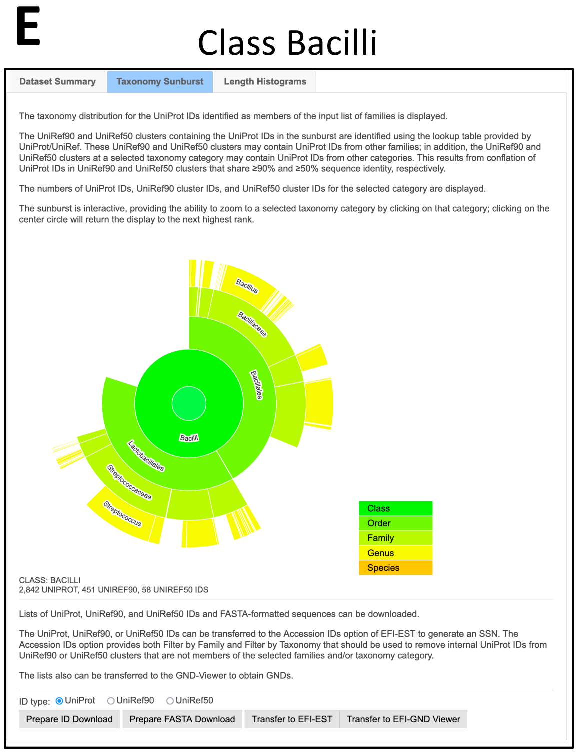
PHYLUM: FIRMICUTES
7,906 UNIPROT, 2,680 UNIREF90, 393 UNIREF50 IDS

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



Class Bacilli

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.

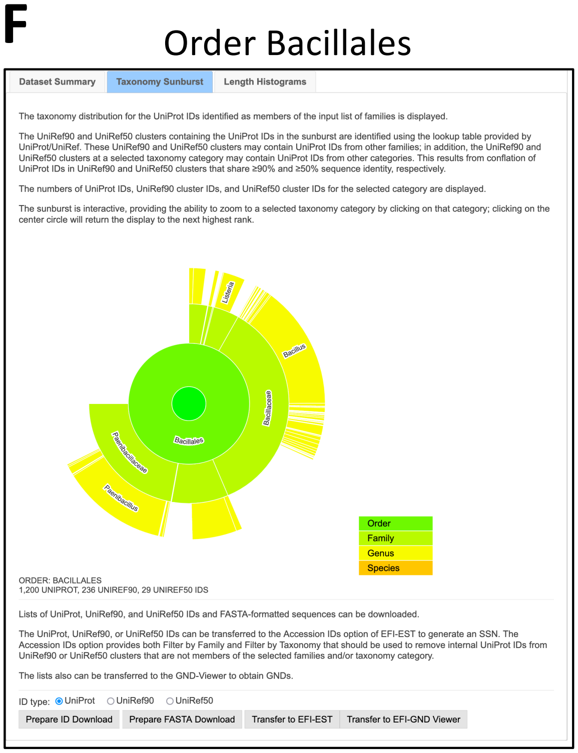
CLASS: BACILLI
2,842 UNIPROT, 451 UNIREF90, 58 UNIREF50 IDS

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



Order Bacillales

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.

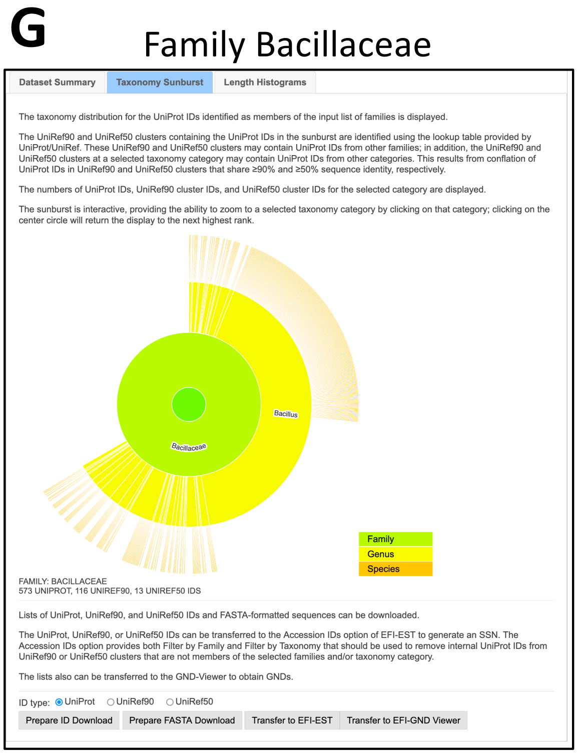
ORDER: BACILLALES
1,200 UNIPROT, 236 UNIREF90, 29 UNIREF50 IDS

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



Family Bacillaceae

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.

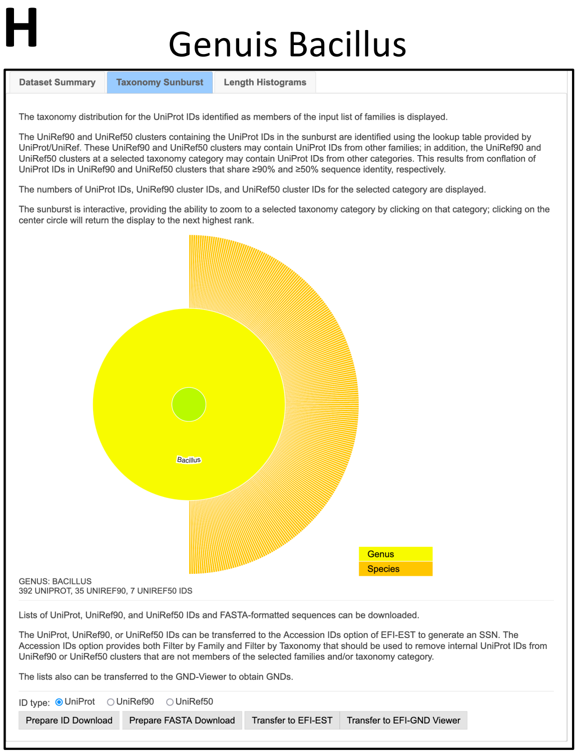
FAMILY: BACILLACEAE
573 UNIPROT, 116 UNIREF90, 13 UNIREF50 IDS

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



Genuis Bacillus

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.

GENUS: BACILLUS
392 UNIPROT, 35 UNIREF90, 7 UNIREF50 IDS

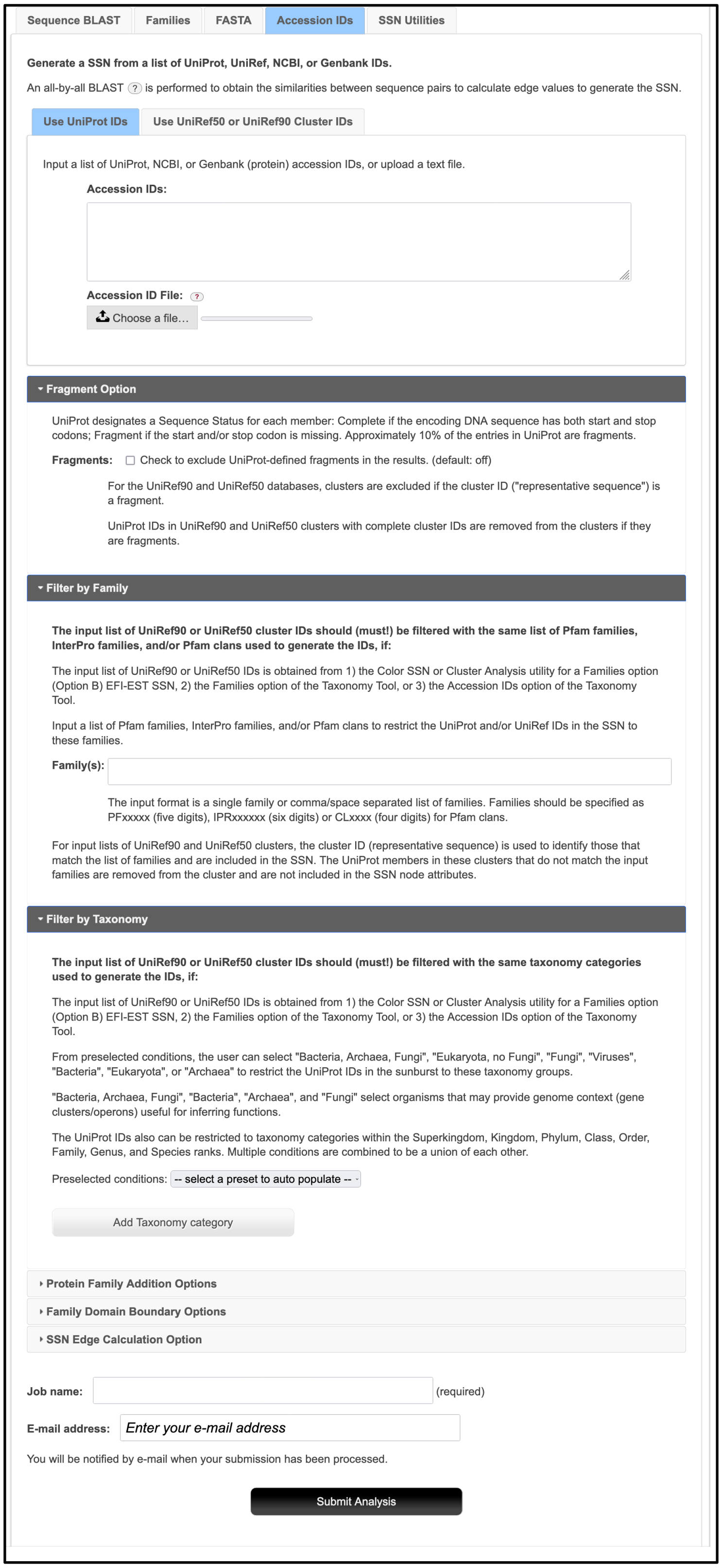
Lists of UniProt, UniRef90, and UniRef50 IDs and FASTA-formatted sequences can be downloaded.

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

Supplementary Figure S4. Selection of Taxonomy Categories for GRE Superfamily in the Taxonomy Sunburst. Panel A, All taxonomy categories (root). **Panel B**, Superkingdom Bacteria. **Panel C**, Kingdom Terrabacteria group. **Panel D**, Phylum Firmicutes. **Panel E**, Class Bacilli. **Panel F**, Order Bacillaes. **Panel G**, Family Bacillaceae. **Panel H**, Genus Bacillus.



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 UniProt, NCBI, or Genbank (protein) accession IDs, or upload a text file.

Accession IDs:

Accession ID File: ?

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 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):

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 --

Protein Family Addition Options

Family Domain Boundary Options

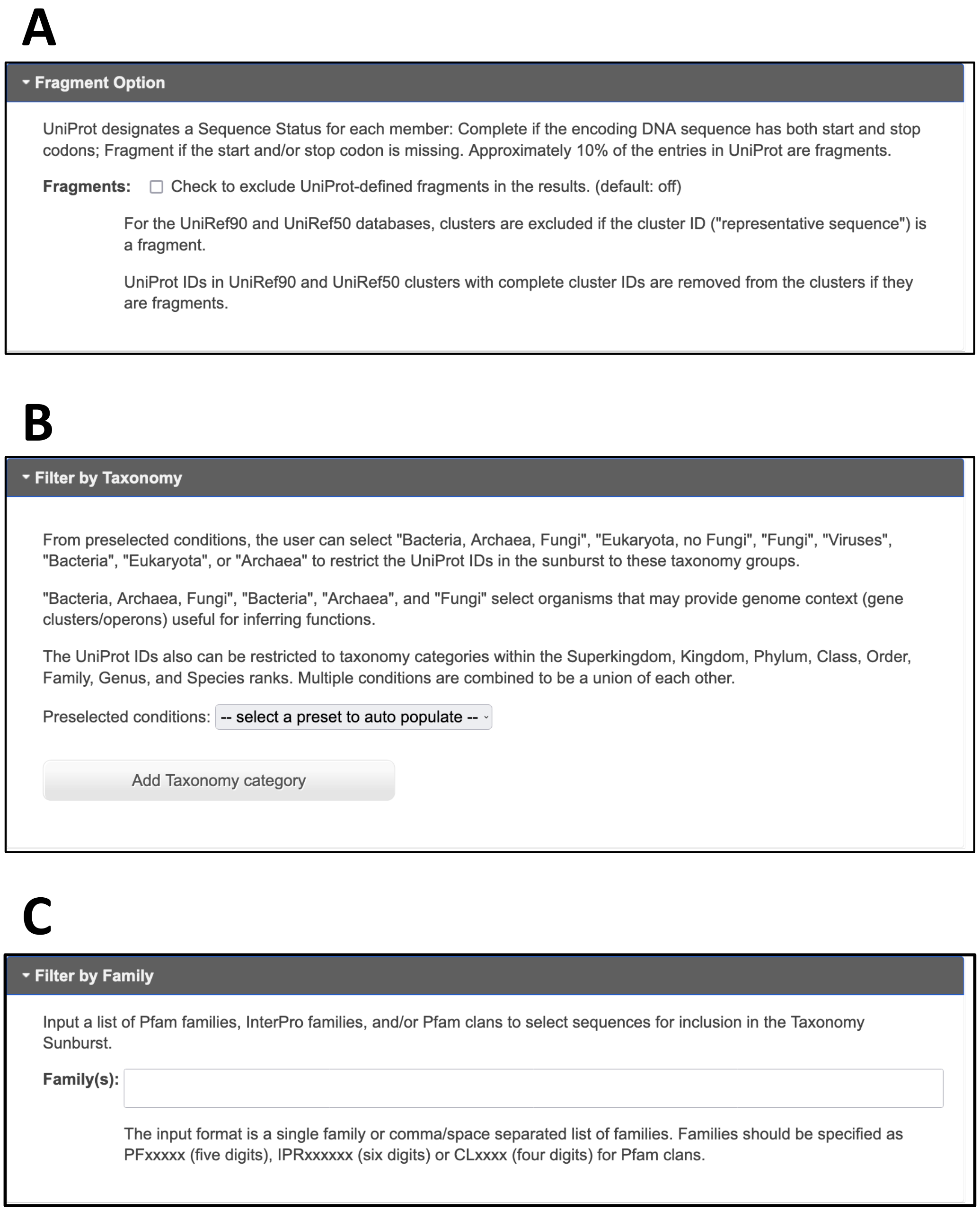
SSN Edge Calculation Option

Job name: (required)

E-mail address: *Enter your e-mail address*

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

Supplementary Figure S5. EFI-EST, Accession IDs Option.



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

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 --

Add Taxonomy category

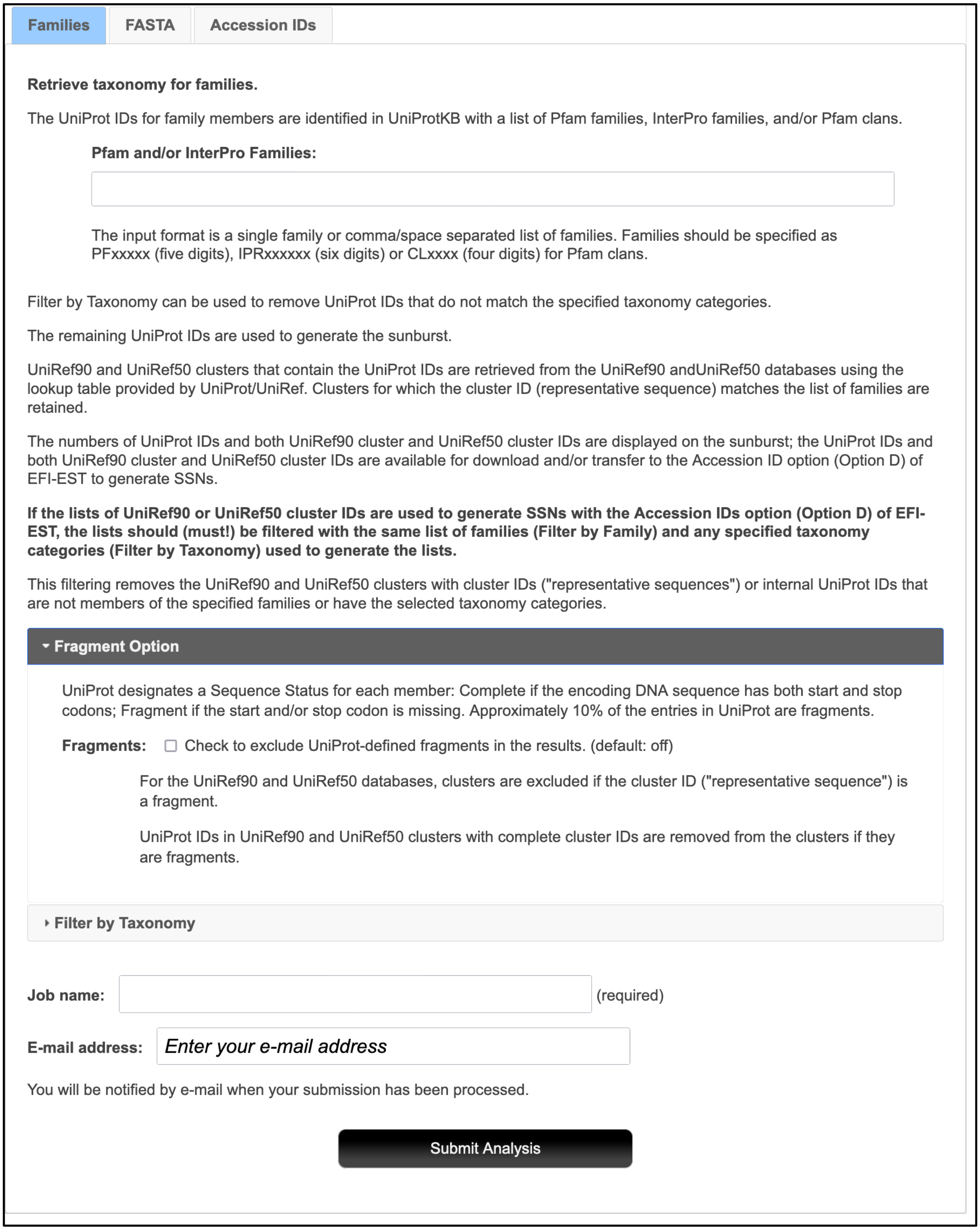
Filter by Family

Input a list of Pfam families, InterPro families, and/or Pfam clans to select sequences for inclusion in the Taxonomy Sunburst.

Family(s):

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

Supplementary Figure S6. Filters. Panel A, Fragment Option. Panel B, Filter by Taxonomy. Panel C, Filter by Family.



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:

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.

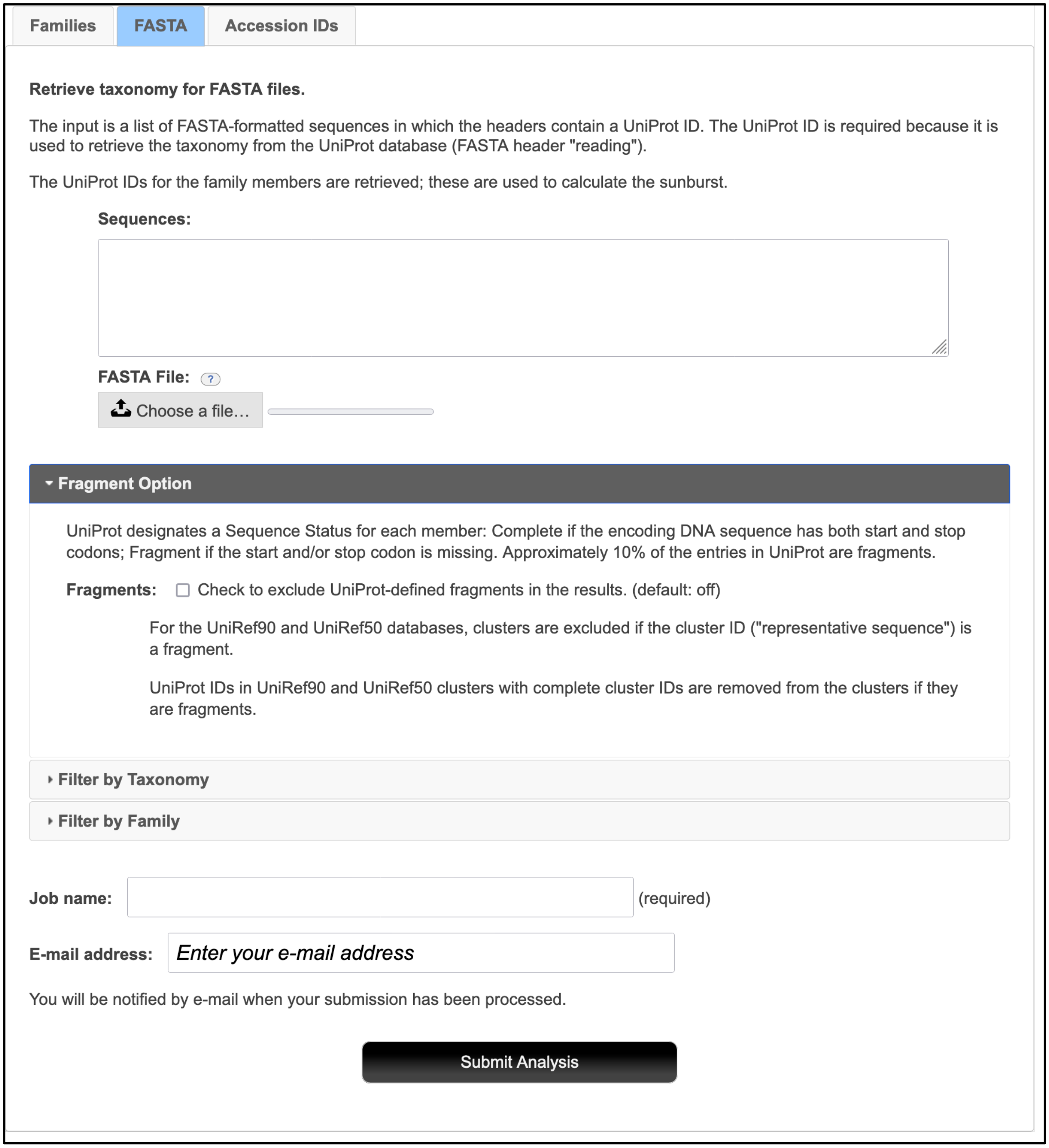
Filter by Taxonomy

Job name: (required)

E-mail address: *Enter your e-mail address*

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

Supplementary Figure S7. Taxonomy Tool, Families Option.



Retrieve taxonomy for FASTA files.

The input is a list of FASTA-formatted sequences in which the headers contain a UniProt ID. The UniProt ID is required because it is used to retrieve the taxonomy from the UniProt database (FASTA header "reading").

The UniProt IDs for the family members are retrieved; these are used to calculate the sunburst.

Sequences:

FASTA File: ?

▾ **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**

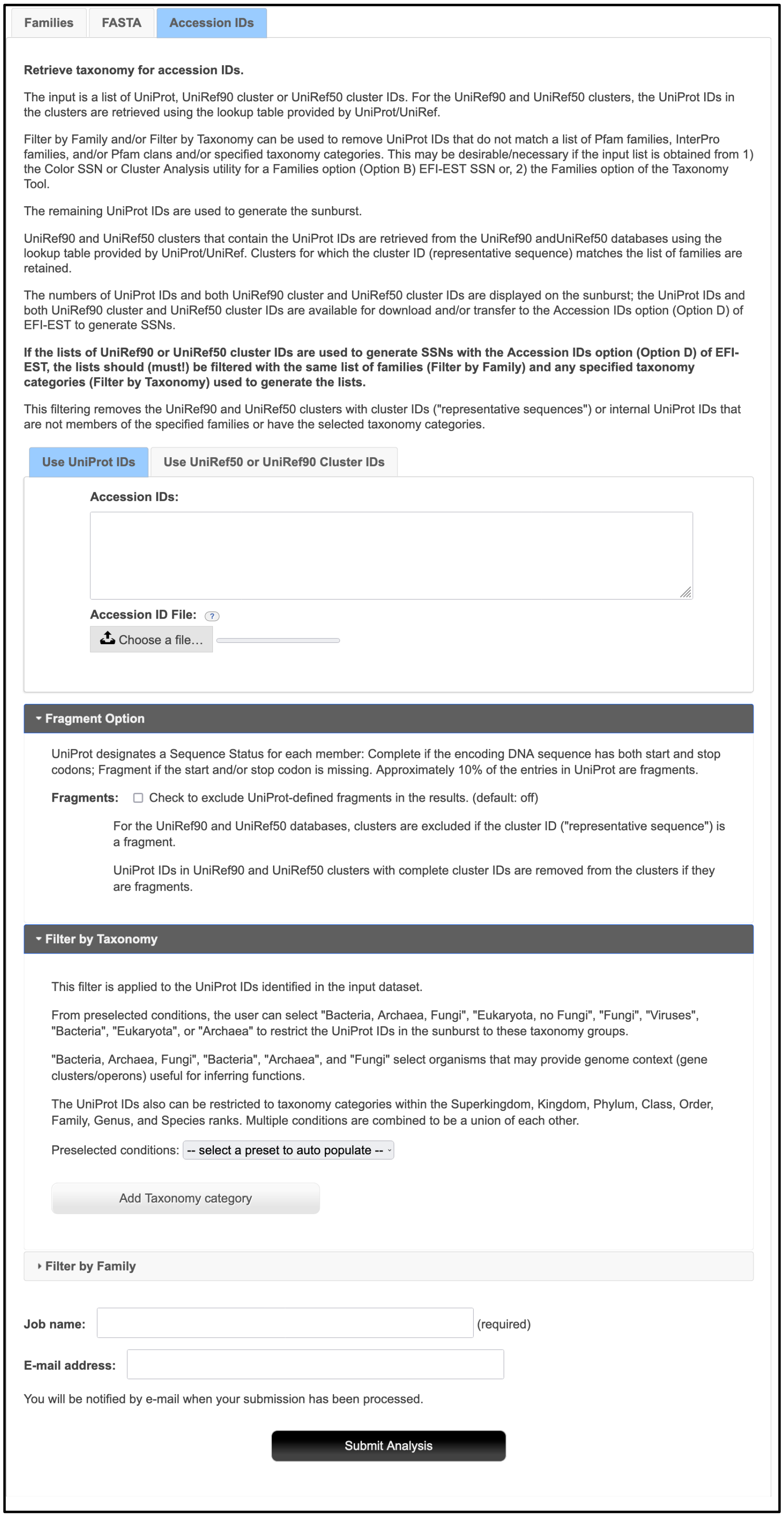
▸ **Filter by Family**

Job name: (required)

E-mail address: *Enter your e-mail address*

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

Supplementary Figure S8. Taxonomy Tool, FASTA Option.



Retrieve taxonomy for accession IDs.

The input is a list of UniProt, UniRef90 cluster or UniRef50 cluster IDs. For the UniRef90 and UniRef50 clusters, the UniProt IDs in the clusters are retrieved using the lookup table provided by UniProt/UniRef.

Filter by Family and/or Filter by Taxonomy can be used to remove UniProt IDs that do not match a list of Pfam families, InterPro families, and/or Pfam clans and/or specified taxonomy categories. This may be desirable/necessary if the input list is obtained from 1) the Color SSN or Cluster Analysis utility for a Families option (Option B) EFI-EST SSN or, 2) the Families option of the Taxonomy Tool.

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 IDs 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.

Accession IDs:

Accession ID File:

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 identified in the input dataset.

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 --

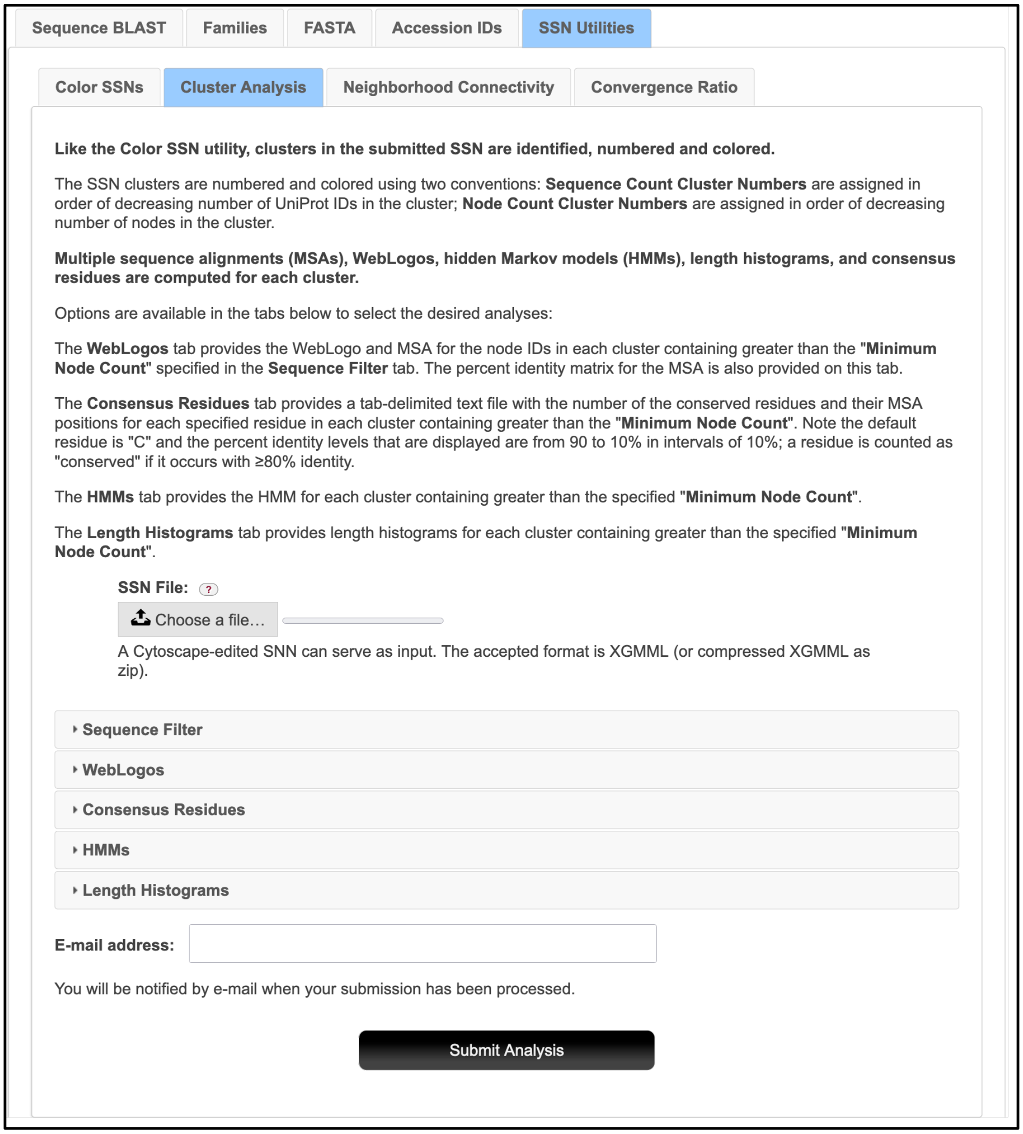
Filter by Family

Job name: (required)

E-mail address:

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

Supplementary Figure S9. Taxonomy Tool, Accession IDs Option.



Like the Color SSN utility, clusters in the submitted SSN are identified, numbered and colored.

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

Multiple sequence alignments (MSAs), WebLogos, hidden Markov models (HMMs), length histograms, and consensus residues are computed for each cluster.

Options are available in the tabs below to select the desired analyses:

The **WebLogos** tab provides the WebLogo and MSA for the node IDs in each cluster containing greater than the "**Minimum Node Count**" specified in the **Sequence Filter** tab. The percent identity matrix for the MSA is also provided on this tab.

The **Consensus Residues** tab provides a tab-delimited text file with the number of the conserved residues and their MSA positions for each specified residue in each cluster containing greater than the "**Minimum Node Count**". Note the default residue is "C" and the percent identity levels that are displayed are from 90 to 10% in intervals of 10%; a residue is counted as "conserved" if it occurs with ≥80% identity.

The **HMMs** tab provides the HMM for each cluster containing greater than the specified "**Minimum Node Count**".

The **Length Histograms** tab provides length histograms for each cluster containing greater than the specified "**Minimum Node Count**".

SSN File: ?

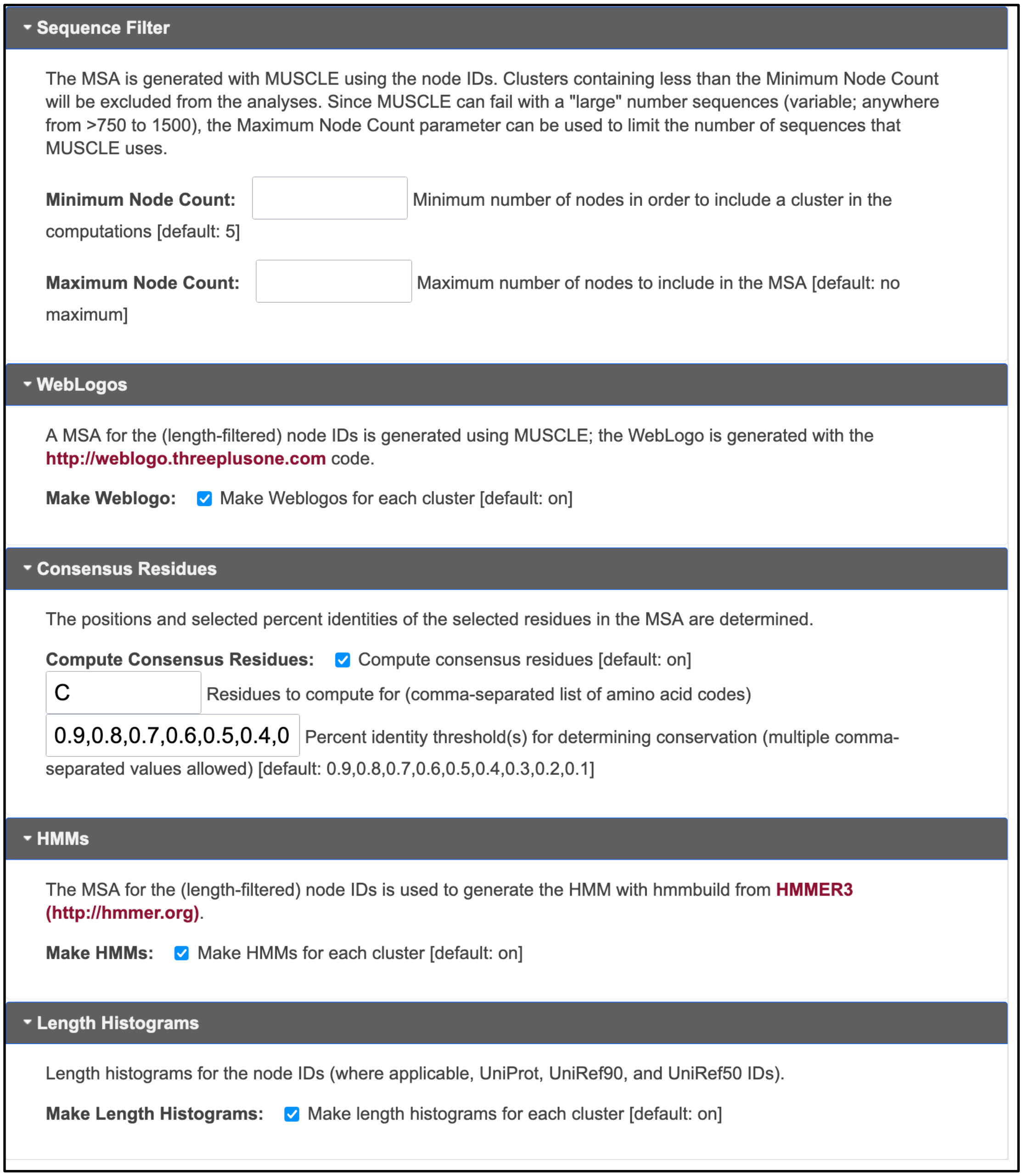
Choose a file…

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.

Supplementary Figure S10. **SSN Utilities: Cluster Analysis Utility.**



▾ Sequence Filter

The MSA is generated with MUSCLE using the node IDs. Clusters containing less than the Minimum Node Count will be excluded from the analyses. Since MUSCLE can fail with a "large" number sequences (variable; anywhere from >750 to 1500), the Maximum Node Count parameter can be used to limit the number of sequences that MUSCLE uses.

Minimum Node Count: ____ Minimum number of nodes in order to include a cluster in the computations [default: 5]

Maximum Node Count: ____ Maximum number of nodes to include in the MSA [default: no maximum]

▾ WebLogos

A MSA for the (length-filtered) node IDs is generated using MUSCLE; the WebLogo is generated with the **http://weblogo.threeplusone.com** code.

Make Weblogo: ☑ Make Weblogos for each cluster [default: on]

▾ Consensus Residues

The positions and selected percent identities of the selected residues in the MSA are determined.

Compute Consensus Residues: ☑ Compute consensus residues [default: on]

C Residues to compute for (comma-separated list of amino acid codes)

0.9,0.8,0.7,0.6,0.5,0.4,0 Percent identity threshold(s) for determining conservation (multiple comma-separated values allowed) [default: 0.9,0.8,0.7,0.6,0.5,0.4,0.3,0.2,0.1]

▾ HMMs

The MSA for the (length-filtered) node IDs is used to generate the HMM with hmmbuild from **HMMER3 (http://hmmer.org)**.

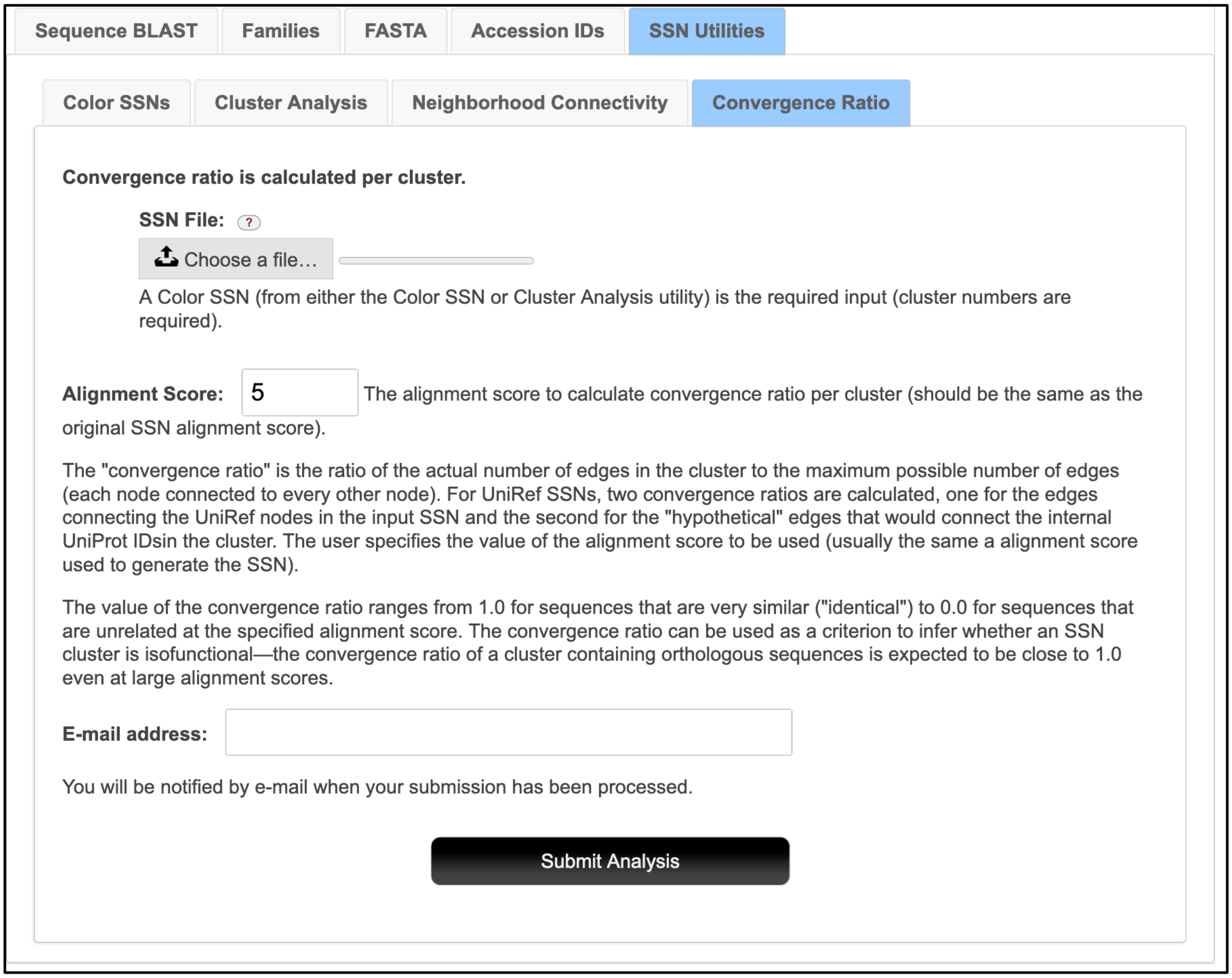
Make HMMs: ☑ Make HMMs for each cluster [default: on]

▾ Length Histograms

Length histograms for the node IDs (where applicable, UniProt, UniRef90, and UniRef50 IDs).

Make Length Histograms: ☑ Make length histograms for each cluster [default: on]

Supplementary Figure S11. SSN Utilities: Cluster Analysis Utility Options.



Convergence ratio is calculated per cluster.

SSN File: ?

A Color SSN (from either the Color SSN or Cluster Analysis utility) is the required input (cluster numbers are required).

Alignment Score: 5 The alignment score to calculate convergence ratio per cluster (should be the same as the original SSN alignment score).

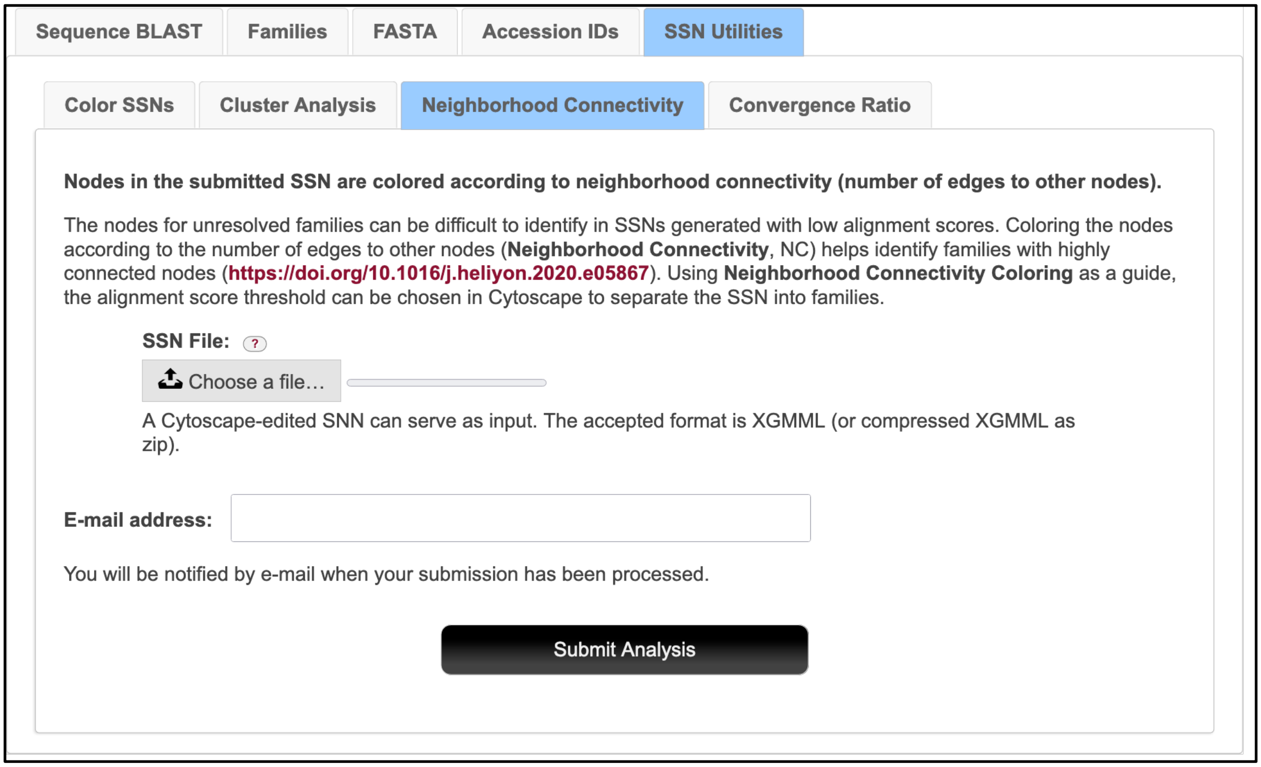
The "convergence ratio" is the ratio of the actual number of edges in the cluster to the maximum possible number of edges (each node connected to every other node). For UniRef SSNs, two convergence ratios are calculated, one for the edges connecting the UniRef nodes in the input SSN and the second for the "hypothetical" edges that would connect the internal UniProt IDsin the cluster. The user specifies the value of the alignment score to be used (usually the same a alignment score used to generate the SSN).

The value of the convergence ratio ranges from 1.0 for sequences that are very similar ("identical") to 0.0 for sequences that are unrelated at the specified alignment score. The convergence ratio can be used as a criterion to infer whether an SSN cluster is isofunctional—the convergence ratio of a cluster containing orthologous sequences is expected to be close to 1.0 even at large alignment scores.

E-mail address:

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

Supplementary Figure S12. SSN Utilities: Convergence Ratio Utility.



Nodes in the submitted SSN are colored according to neighborhood connectivity (number of edges to other nodes).

The nodes for unresolved families can be difficult to identify in SSNs generated with low alignment scores. Coloring the nodes according to the number of edges to other nodes (**Neighborhood Connectivity**, NC) helps identify families with highly connected nodes (**https://doi.org/10.1016/j.heliyon.2020.e05867**). Using **Neighborhood Connectivity Coloring** as a guide, the alignment score threshold can be chosen in Cytoscape to separate the SSN into families.

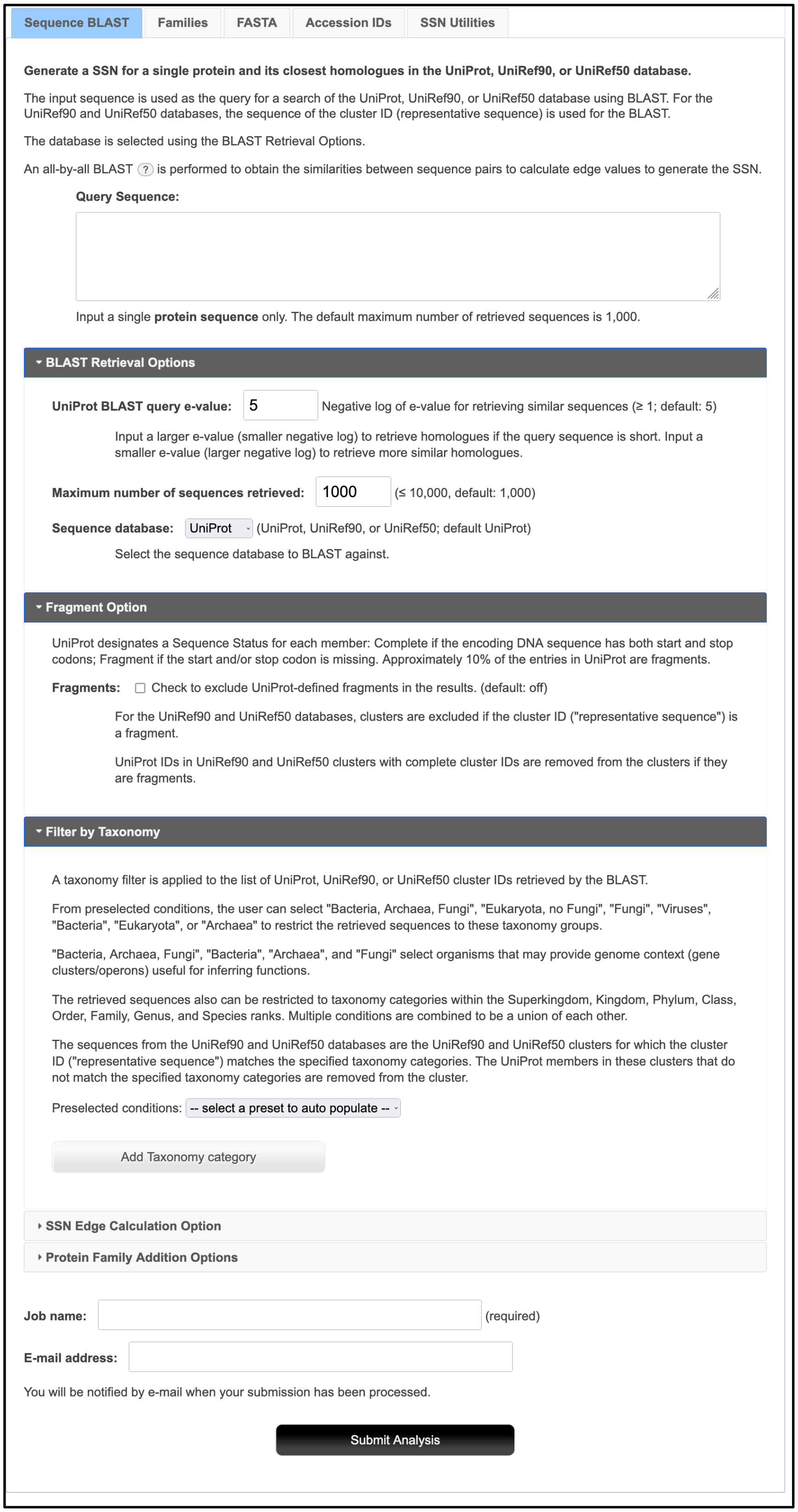
SSN File: ?

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.

Supplementary Figure S13. **SSN Utilities: Neighborhood Connectivity Utility.**



Generate a SSN for a single protein and its closest homologues in the UniProt, UniRef90, or UniRef50 database.

The input sequence is used as the query for a search of the UniProt, UniRef90, or UniRef50 database using BLAST. For the UniRef90 and UniRef50 databases, the sequence of the cluster ID (representative sequence) is used for the BLAST.

The database is selected using the BLAST Retrieval Options.

An all-by-all BLAST ? is performed to obtain the similarities between sequence pairs to calculate edge values to generate the SSN.

Query Sequence:

Input a single **protein sequence** only. The default maximum number of retrieved sequences is 1,000.

BLAST Retrieval Options

UniProt BLAST query e-value: 5 Negative log of e-value for retrieving similar sequences (≥ 1; default: 5)

Input a larger e-value (smaller negative log) to retrieve homologues if the query sequence is short. Input a smaller e-value (larger negative log) to retrieve more similar homologues.

Maximum number of sequences retrieved: 1000 (≤ 10,000, default: 1,000)

Sequence database: UniProt (UniProt, UniRef90, or UniRef50; default UniProt)

Select the sequence database to BLAST against.

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

A taxonomy filter is applied to the list of UniProt, UniRef90, or UniRef50 cluster IDs retrieved by the BLAST.

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 --

SSN Edge Calculation Option

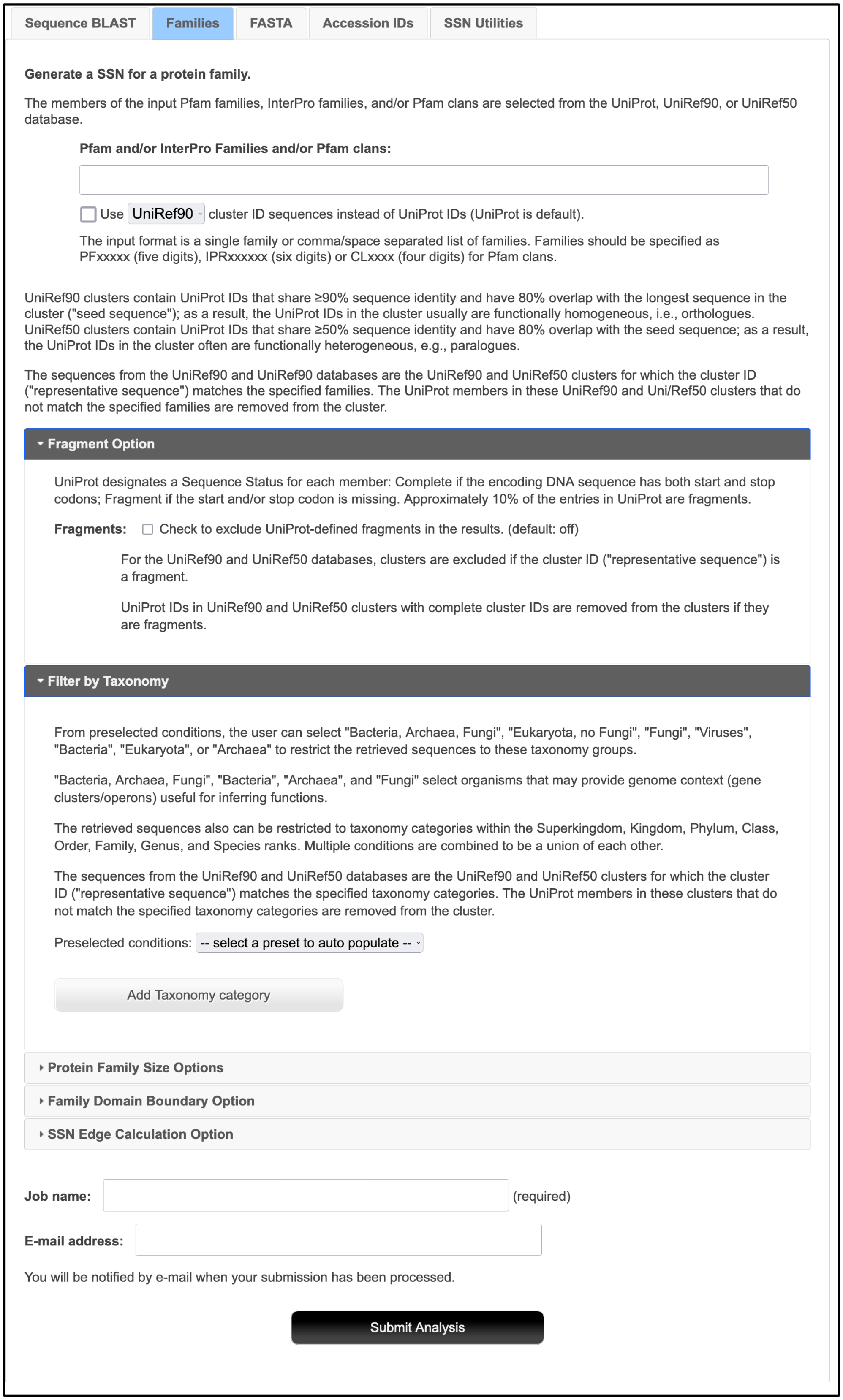
Protein Family Addition Options

Job name: (required)

E-mail address:

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

Supplementary Figure S14. EFI-EST, Sequence BLAST Option.



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:

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

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 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

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

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

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

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

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

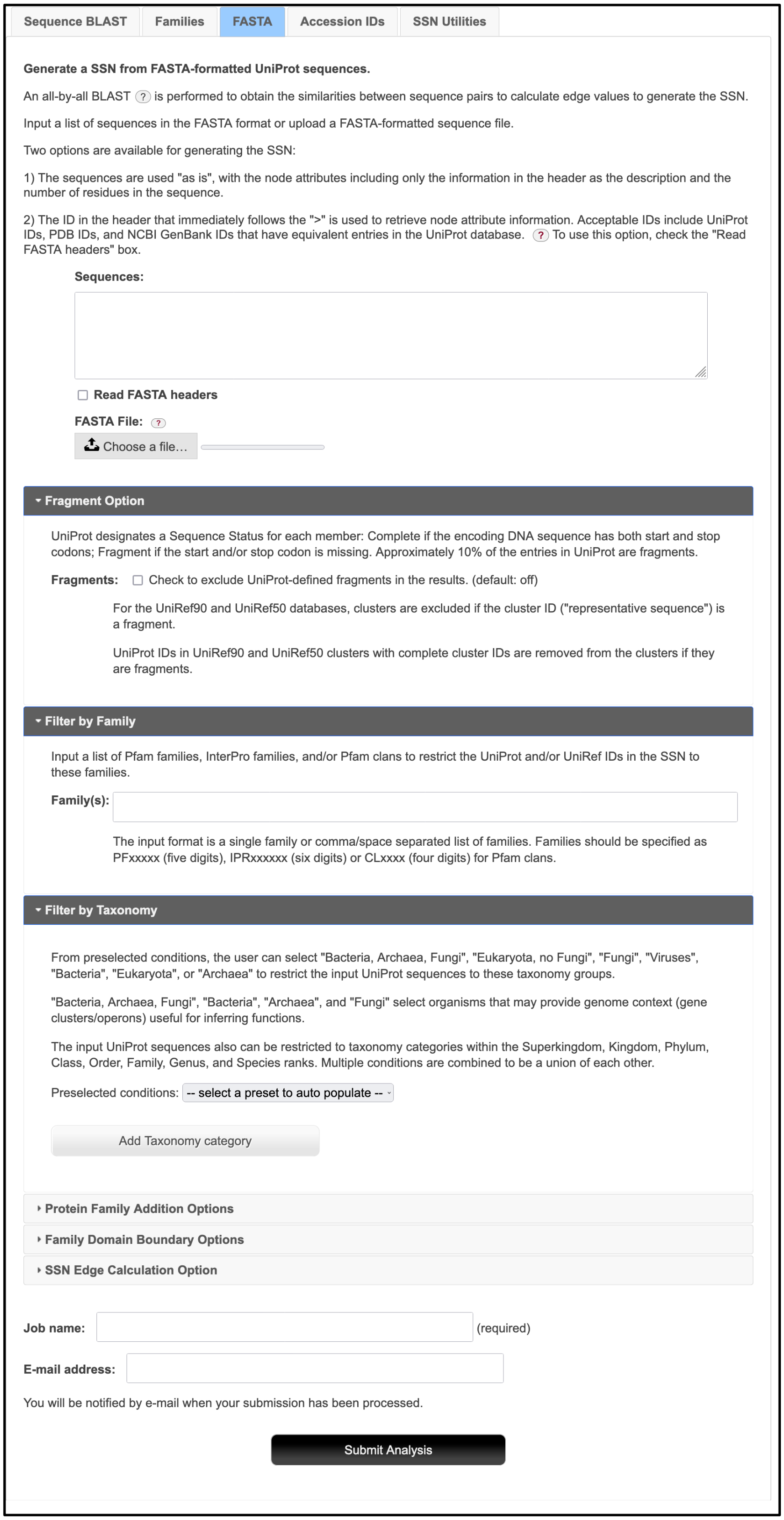
Add Taxonomy category
Protein Family Size Options
Family Domain Boundary Option
SSN Edge Calculation Option

Job name: (required)

E-mail address:

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

Supplementary Figure S15. **EFI-EST, Families Option.**



Generate a SSN from FASTA-formatted UniProt sequences.

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 sequences in the FASTA format or upload a FASTA-formatted sequence file.

Two options are available for generating the SSN:

1) The sequences are used "as is", with the node attributes including only the information in the header as the description and the number of residues in the sequence.

2) The ID in the header that immediately follows the ">" is used to retrieve node attribute information. Acceptable IDs include UniProt IDs, PDB IDs, and NCBI GenBank IDs that have equivalent entries in the UniProt database. ? To use this option, check the "Read FASTA headers" box.

Sequences:

□ **Read FASTA headers**

FASTA File: ?

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.

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

Family(s):

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

From preselected conditions, the user can select "Bacteria, Archaea, Fungi", "Eukaryota, no Fungi", "Fungi", "Viruses", "Bacteria", "Eukaryota", or "Archaea" to restrict the input UniProt 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 input UniProt 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.

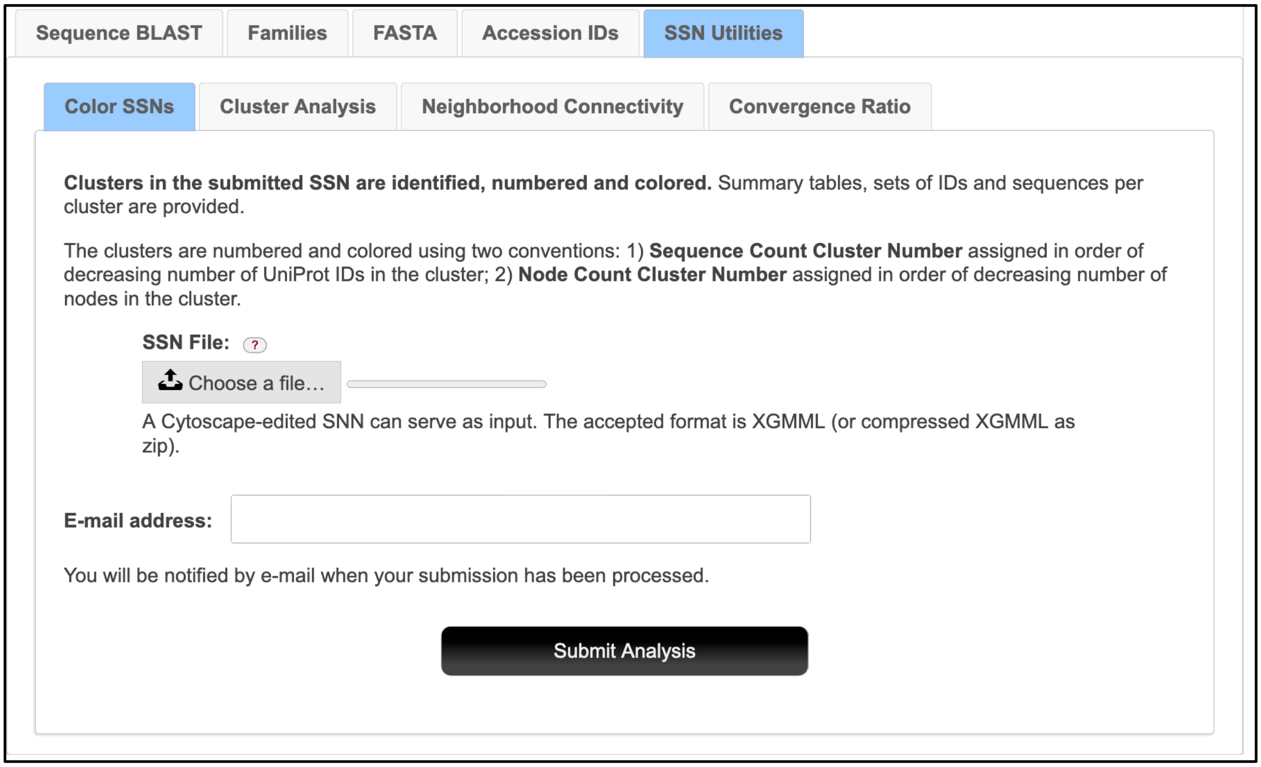
Preselected conditions: -- select a preset to auto populate --

Job name: (required)

E-mail address:

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

Supplementary Figure S16. EFI-EST, FASTA Option.



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: ?

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.

Supplementary Figure S17. **SSN Utilities: Color SSNs.**