Supplementary Methods

The EFI Web Resource for Genomic Enzymology Tools: Leveraging Protein, Genome, and Metagenome Databases to Discover Novel Enzymes and Metabolic Pathways

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

[1]Institute for Genomic Biology, [2]Department of Biochemistry, and [3]Department of Chemistry, University of Illinois at Urbana-Champaign, 1206 West Gregory Drive, Urbana, Illinois 61801, United States.

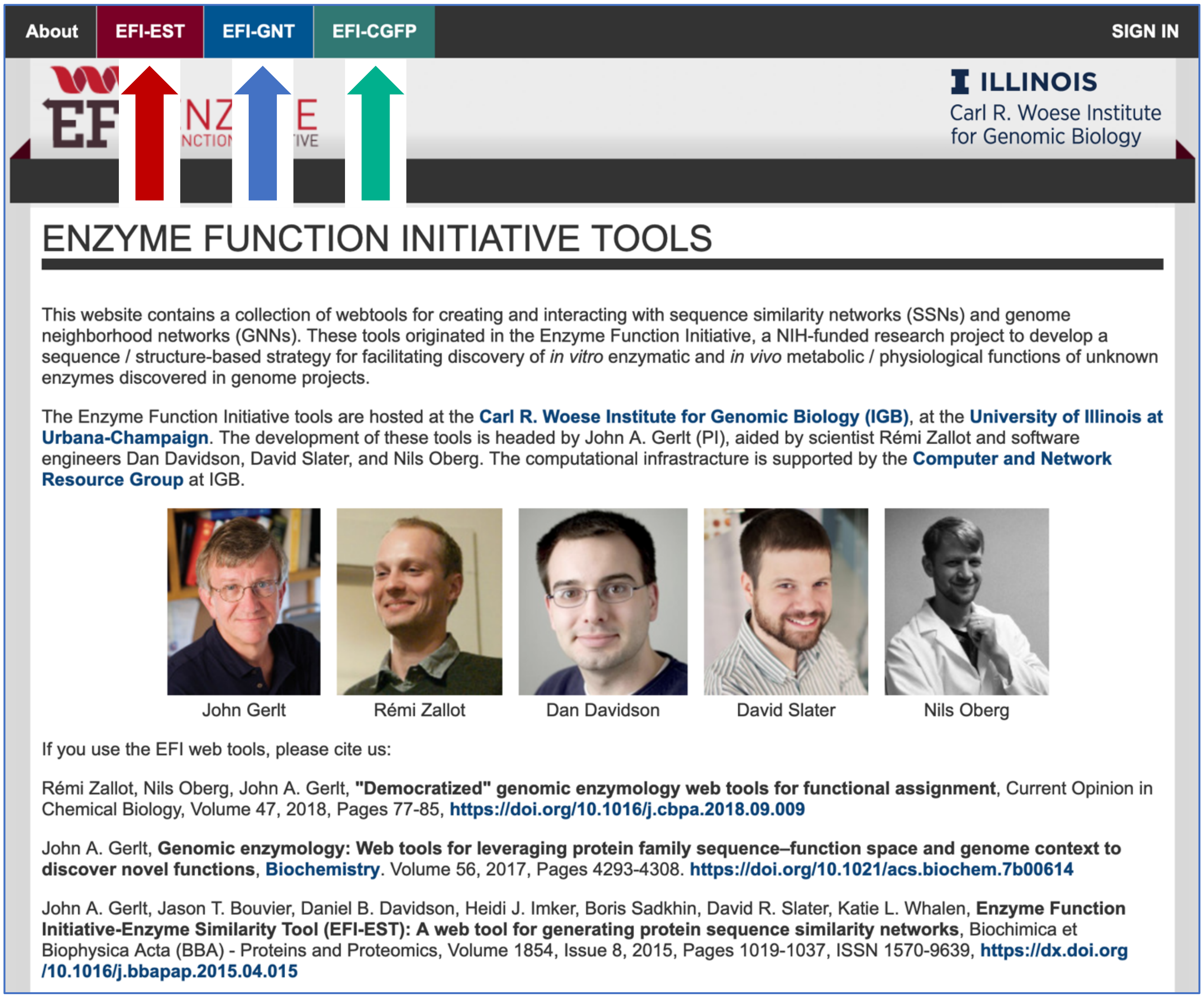
Corresponding author: John A. Gerlt (j-gerlt@illinois.edu).

Phone: +1-217-244-7414

TABLE OF CONTENTS

Introduction

The EFI-EST, EFI-GNT, and EFI-CGFP web tools are accessed by tabs (colored arrows) on the Enzyme Function Initiative Tools home page (https://efi.igb.illinois.edu/index.php) as well as the pages for each tool.



ENZYME FUNCTION INITIATIVE TOOLS

This website contains a collection of webtools for creating and interacting with sequence similarity networks (SSNs) and genome neighborhood networks (GNNs). These tools originated in the Enzyme Function Initiative, a NIH-funded research project to develop a sequence / structure-based strategy for facilitating discovery of *in vitro* enzymatic and *in vivo* metabolic / physiological functions of unknown enzymes discovered in genome projects.

The Enzyme Function Initiative tools are hosted at the **Carl R. Woese Institute for Genomic Biology (IGB)**, at the **University of Illinois at Urbana-Champaign**. The development of these tools is headed by John A. Gerlt (PI), aided by scientist Rémi Zallot and software engineers Dan Davidson, David Slater, and Nils Oberg. The computational infrastructure is supported by the **Computer and Network Resource Group** at IGB.

If you use the EFI web tools, please cite us:

Rémi Zallot, Nils Oberg, John A. Gerlt, **"Democratized" genomic enzymology web tools for functional assignment**, Current Opinion in Chemical Biology, Volume 47, 2018, Pages 77-85, **https://doi.org/10.1016/j.cbpa.2018.09.009**

John A. Gerlt, **Genomic enzymology: Web tools for leveraging protein family sequence–function space and genome context to discover novel functions**, **Biochemistry**. Volume 56, 2017, Pages 4293-4308. **https://doi.org/10.1021/acs.biochem.7b00614**

John A. Gerlt, Jason T. Bouvier, Daniel B. Davidson, Heidi J. Imker, Boris Sadkhin, David R. Slater, Katie L. Whalen, **Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): A web tool for generating protein sequence similarity networks**, Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics, Volume 1854, Issue 8, 2015, Pages 1019-1037, ISSN 1570-9639, **https://dx.doi.org/10.1016/j.bbapap.2015.04.015**

These On-Line Methods provide detailed descriptions of the tools, including screen shots of the pages for submitting jobs and downloading files; tables also are included that describe the node attributes that accompany the SSNs generated by the tools. These Methods are intended to serve as a “user’s manual”, instructing the user on the operation of the tools, anticipating questions, and providing advice on the selection of input parameters.

I. Background Information

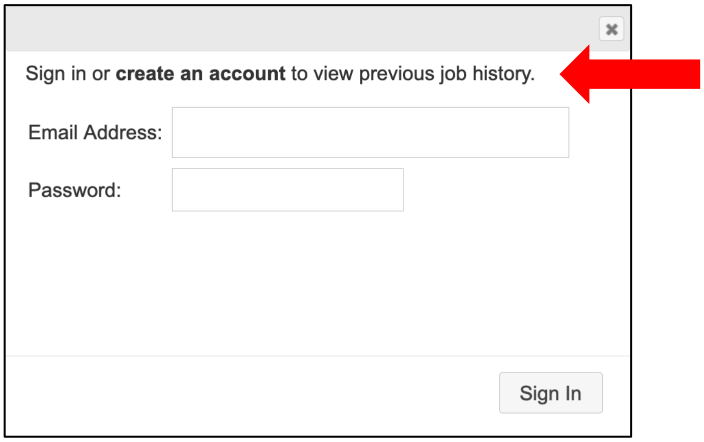
The tools facilitate investigation of sequence-function relationships in protein families curated by the Pfam (https://pfam.xfam.org/; currently Release 32.0, with 17,929 sequence-based domain families and 628 clans) and InterPro databases (https://www.ebi.ac.uk/interpro/; currently Release 74, with 35,484 structure- and sequence-based homologous superfamilies, domains, and families defined by its fourteen consortium member databases, including Pfam). Both Pfam and InterPro use protein sequences from UniProtKB and genome sequences from the European Nucleotide Archive (ENA; https://www.ebi.ac.uk/ena). Pfam and InterPro families were selected for the tools because annotations in the UniProt database can be corrected by community input (https://www.uniprot.org/update); the National Center for Biotechnology Information (NCBI) maintains (larger) protein and genome databases, archives that can be updated/corrected only by depositors.

The tools use sequence similarity networks (SSNs) to visualize/analyze sequence-function space in protein families as well as transfer information among the tools. Briefly, an SSN is a multi-dimensional network displaying the pairwise sequence similarity relationships among proteins[1]. Each protein is represented by a symbol (“node”); two nodes are connected by a line (“edge”) if they share pairwise sequence similarity that exceeds a specified threshold. The use of

SSNs to explore sequence-function space in proteins families has been described[2-5]; users are referred to those publications for background information about SSNs.

A. User Accounts

User accounts are optional for EFI-EST and EFI-GNT; they are required for EFI-CGFP (*vide infra*). An account is created by clicking the "**create an account**" button (red arrow) on the form that is provided when the "**SIGN IN**" button at the top left of the home page is clicked:

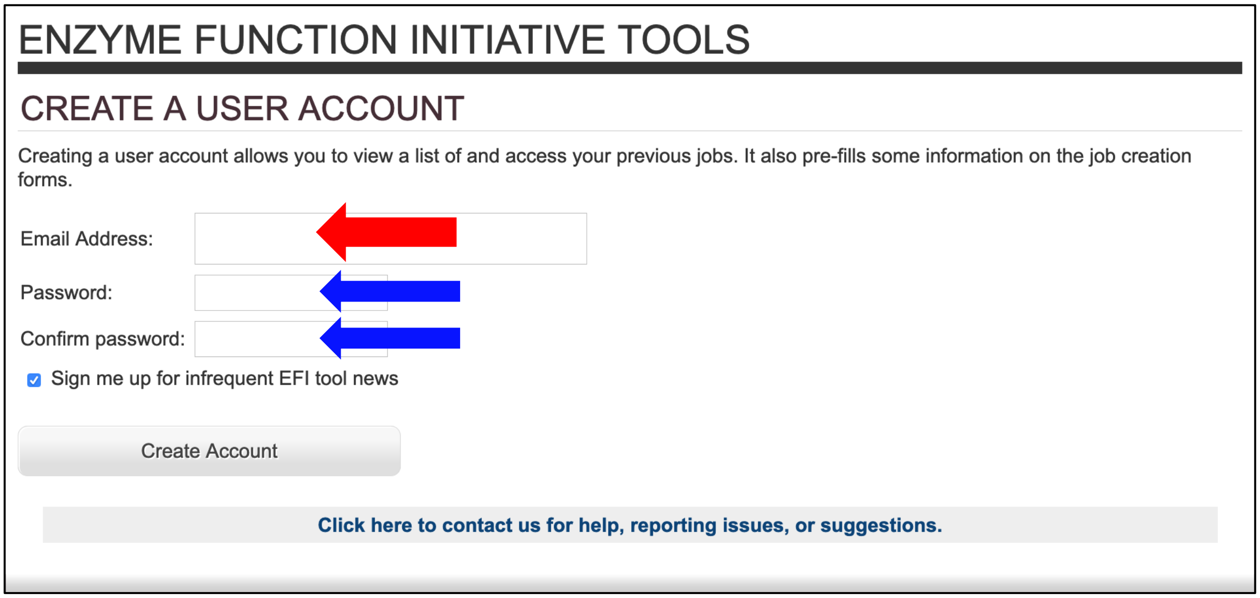


Sign in or **create an account** to view previous job history.

Email Address:

Password:

An e-mail address is entered (red arrow); a password is entered and confirmed (blue arrows):



ENZYME FUNCTION INITIATIVE TOOLS

CREATE A USER ACCOUNT

Creating a user account allows you to view a list of and access your previous jobs. It also pre-fills some information on the job creation forms.

Email Address:

Password:

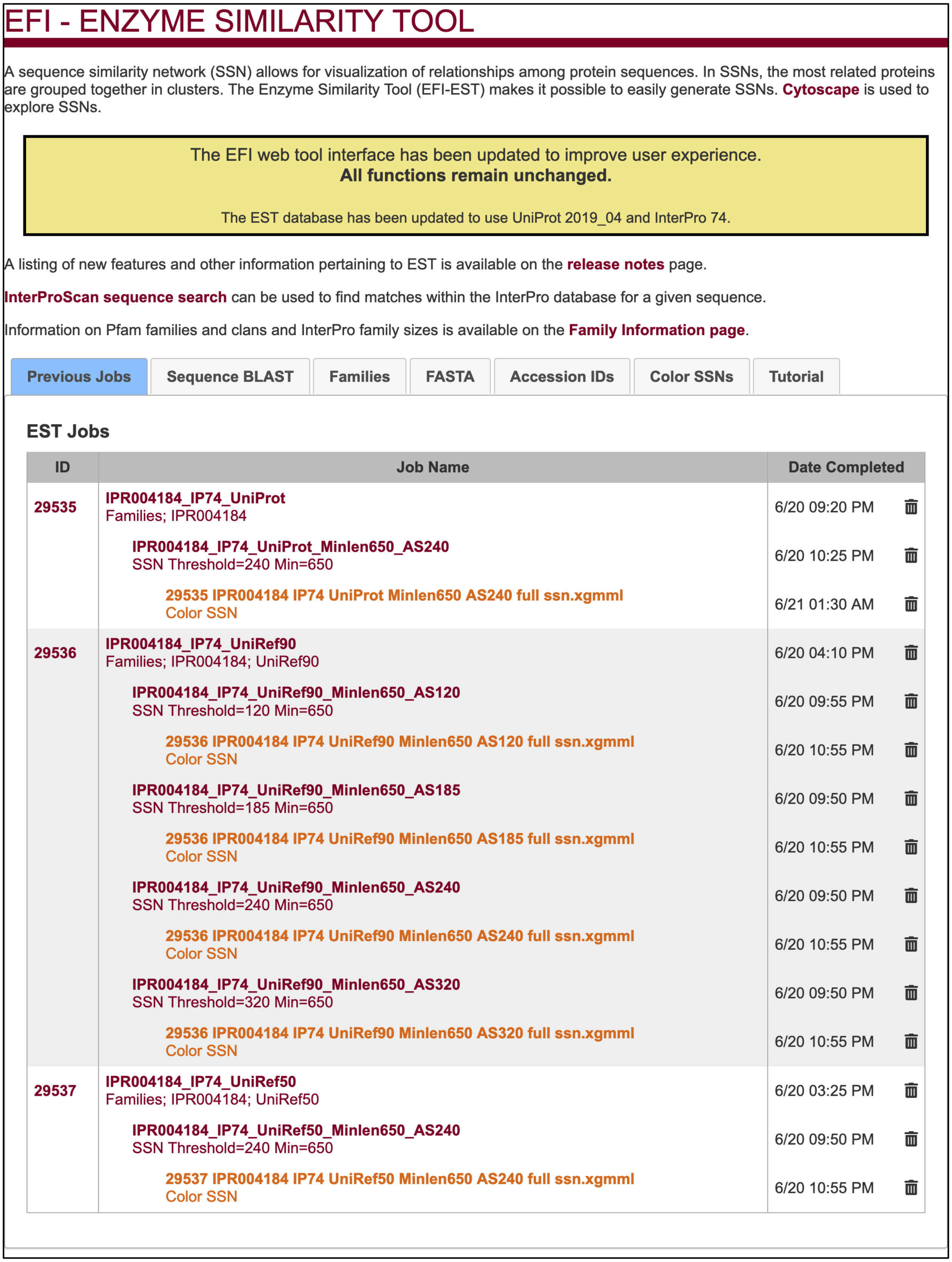
Confirm password:

☑ Sign me up for infrequent EFI tool news

Create Account
Click here to contact us for help, reporting issues, or suggestions.

When the account is activated (by clicking on a link in an e-mail verification request), the user enters his/her e-mail and password with each "**SIGN IN**".

With an account, the home page for each tool includes a "**Previous Jobs**" panel that lists the identities and status of active jobs ("PENDING" or "RUNNING"), the identities of failed jobs ("FAILED"), and links to the output pages for all jobs completed within the previous 40 days, making it easier to assess the status of submitted jobs and locate the results of completed jobs. The following panel summarizes the EFI-EST jobs performed to provide SSNs provided in the main text for the analyses of the glycyl radical enzyme (GRE) superfamily (IPR004184).



EFI - ENZYME SIMILARITY TOOL

A sequence similarity network (SSN) allows for visualization of relationships among protein sequences. In SSNs, the most related proteins are grouped together in clusters. The Enzyme Similarity Tool (EFI-EST) makes it possible to easily generate SSNs. **Cytoscape** is used to explore SSNs.

The EFI web tool interface has been updated to improve user experience.
All functions remain unchanged.

The EST database has been updated to use UniProt 2019_04 and InterPro 74.

A listing of new features and other information pertaining to EST is available on the **release notes** page.

InterProScan sequence search can be used to find matches within the InterPro database for a given sequence.

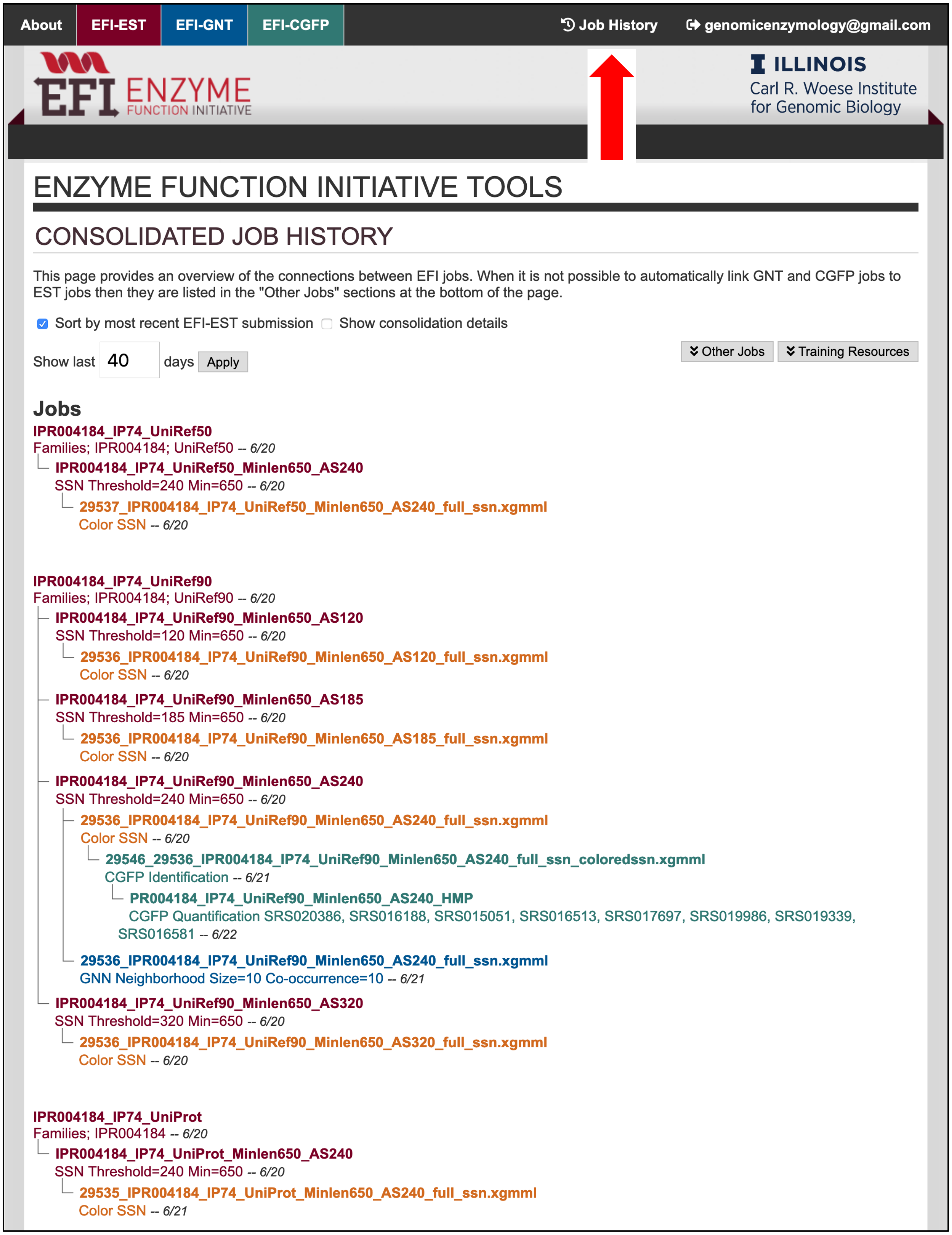
Information on Pfam families and clans and InterPro family sizes is available on the **Family Information page**.

EST Jobs

ID	Job Name	Date Completed
29535	**IPR004184_IP74_UniProt** Families; IPR004184	6/20 09:20 PM 🗑
	IPR004184_IP74_UniProt_Minlen650_AS240 SSN Threshold=240 Min=650	6/20 10:25 PM 🗑
	29535 IPR004184 IP74 UniProt Minlen650 AS240 full ssn.xgmml Color SSN	6/21 01:30 AM 🗑
29536	**IPR004184_IP74_UniRef90** Families; IPR004184; UniRef90	6/20 04:10 PM 🗑
	IPR004184_IP74_UniRef90_Minlen650_AS120 SSN Threshold=120 Min=650	6/20 09:55 PM 🗑
	29536 IPR004184 IP74 UniRef90 Minlen650 AS120 full ssn.xgmml Color SSN	6/20 10:55 PM 🗑
	IPR004184_IP74_UniRef90_Minlen650_AS185 SSN Threshold=185 Min=650	6/20 09:50 PM 🗑
	29536 IPR004184 IP74 UniRef90 Minlen650 AS185 full ssn.xgmml Color SSN	6/20 10:55 PM 🗑
	IPR004184_IP74_UniRef90_Minlen650_AS240 SSN Threshold=240 Min=650	6/20 09:50 PM 🗑
	29536 IPR004184 IP74 UniRef90 Minlen650 AS240 full ssn.xgmml Color SSN	6/20 10:55 PM 🗑
	IPR004184_IP74_UniRef90_Minlen650_AS320 SSN Threshold=320 Min=650	6/20 09:50 PM 🗑
	29536 IPR004184 IP74 UniRef90 Minlen650 AS320 full ssn.xgmml Color SSN	6/20 10:55 PM 🗑
29537	**IPR004184_IP74_UniRef50** Families; IPR004184; UniRef50	6/20 03:25 PM 🗑
	IPR004184_IP74_UniRef50_Minlen650_AS240 SSN Threshold=240 Min=650	6/20 09:50 PM 🗑
	29537 IPR004184 IP74 UniRef50 Minlen650 AS240 full ssn.xgmml Color SSN	6/20 10:55 PM 🗑

The tools also send e-mails to the user when jobs are started and finished, with the latter providing links to the results pages; however, these are easy to lose in e-mail in-boxes. Links to completed or failed jobs can be removed from the "**Previous Jobs**" panel with the Trash icon ("**Archive**"). Running jobs can be stopped and removed from the "**Previous Jobs**" panel using the Stop icon.

A user account also provides access to the "**Job History**" page (accessed at the top of each page; red arrow).


About
EFI-EST
EFI-GNT
EFI-CGFP
Job History
genomicenzymology@gmail.com
EFI ENZYME FUNCTION INITIATIVE
ILLINOIS
Carl R. Woese Institute for Genomic Biology
Other Jobs
Training Resources
Apply

ENZYME FUNCTION INITIATIVE TOOLS

CONSOLIDATED JOB HISTORY

This page provides an overview of the connections between EFI jobs. When it is not possible to automatically link GNT and CGFP jobs to EST jobs then they are listed in the "Other Jobs" sections at the bottom of the page.

☑ Sort by most recent EFI-EST submission □ Show consolidation details

Show last 40 days

Jobs

```
IPR004184_IP74_UniRef50
Families; IPR004184; UniRef50 -- 6/20
└ IPR004184_IP74_UniRef50_Minlen650_AS240
  SSN Threshold=240 Min=650 -- 6/20
  └ 29537_IPR004184_IP74_UniRef50_Minlen650_AS240_full_ssn.xgmml
    Color SSN -- 6/20

IPR004184_IP74_UniRef90
Families; IPR004184; UniRef90 -- 6/20
├ IPR004184_IP74_UniRef90_Minlen650_AS120
│ SSN Threshold=120 Min=650 -- 6/20
│ └ 29536_IPR004184_IP74_UniRef90_Minlen650_AS120_full_ssn.xgmml
│   Color SSN -- 6/20
├ IPR004184_IP74_UniRef90_Minlen650_AS185
│ SSN Threshold=185 Min=650 -- 6/20
│ └ 29536_IPR004184_IP74_UniRef90_Minlen650_AS185_full_ssn.xgmml
│   Color SSN -- 6/20
├ IPR004184_IP74_UniRef90_Minlen650_AS240
│ SSN Threshold=240 Min=650 -- 6/20
│ ├ 29536_IPR004184_IP74_UniRef90_Minlen650_AS240_full_ssn.xgmml
│ │ Color SSN -- 6/20
│ │ └ 29546_29536_IPR004184_IP74_UniRef90_Minlen650_AS240_full_ssn_coloredssn.xgmml
│ │   CGFP Identification -- 6/21
│ │   └ PR004184_IP74_UniRef90_Minlen650_AS240_HMP
│ │     CGFP Quantification SRS020386, SRS016188, SRS015051, SRS016513, SRS017697, SRS019986, SRS019339,
│ │   SRS016581 -- 6/22
│ └ 29536_IPR004184_IP74_UniRef90_Minlen650_AS240_full_ssn.xgmml
│   GNN Neighborhood Size=10 Co-occurrence=10 -- 6/21
└ IPR004184_IP74_UniRef90_Minlen650_AS320
  SSN Threshold=320 Min=650 -- 6/20
  └ 29536_IPR004184_IP74_UniRef90_Minlen650_AS320_full_ssn.xgmml
    Color SSN -- 6/20

IPR004184_IP74_UniProt
Families; IPR004184 -- 6/20
└ IPR004184_IP74_UniProt_Minlen650_AS240
  SSN Threshold=240 Min=650 -- 6/20
  └ 29535_IPR004184_IP74_UniProt_Minlen650_AS240_full_ssn.xgmml
    Color SSN -- 6/21
```

This page provides a “parent-child tree” summary of the succession of EFI-EST, EFI-GNT and EFI-CGFP jobs generated from each initial EFI-EST BLAST job, i.e., the “**Job History**” integrates the entries on the “**Previous Jobs**” panels on the user’s home pages for EFI-EST, EFI-GNT, and EFI-CGFP. Each entry is a link to a job so the user can quickly access each of the analyses that were performed. This “**Job History**” page summarizes the jobs that were described in the main text for the analyses of the glycyl radical enzyme (GRE) superfamily (IPR004184).

Users must be pre-approved to use EFI-CGFP—the computational requirements are significant, so a description of the anticipated use is required before access is granted.

The next sections discuss several important topics that are pertinent to the use of all three tools: alignment scores for edges in SSNs[1-4, 6], visualization of SSNs using Cytoscape, the BridgeDB Cytoscape app for adding “custom” node attributes to SSNs, SSNs for large protein families, and SSNs for multidomain proteins. Users are advised to read these sections before using the tools. The tools then are described.

B. Alignment Scores for Edges

A minimum sequence similarity threshold for drawing edges between SSN nodes needs to be chosen with the goal of segregating nodes into “isofunctional” clusters. The values of the edges are derived from pairwise comparisons of the sequences in the input dataset using BLAST. BLAST provides several measures of sequence similarity, including the global percent identity between query and subject sequences as well as statistical scores that describe the quality of the alignment (including the presence/importance of gaps when divergent sequences are aligned).

The “e-value”, familiar to most users, is the *database size-dependent* probability that the sequence similarity between the query and subject sequences occurs by chance; it is calculated from the “bit score” that provides a *database size-independent* measure of the quality of the alignment. The magnitude of the e-value can be interpreted *qualitatively* in terms of sequence similarity, e.g., the smaller the e-value (the larger its negative logarithm), the greater the pairwise similarity (less likely that the sequence similarity occurs by chance). However, the magnitude of the e-value is dependent on the lengths of the query and subject sequences, the presence and size of gaps when the sequences are aligned, and the number of sequences in the database; therefore, its magnitude *is not* a quantitative measure of sequence similarity.

To partially address these complications, EFI-EST assigns a *database size-independent* “alignment score” to each edge that is calculated from the bit-score. The alignment score is $-\log_{10}[2^{-B}(L_Q L_S)]$, where L_Q is the query length, L_S is the subject length, and B is the bitscore.

The alignment score is similar in magnitude to the negative base-10 logarithm of the e-value, so the alignment score also can be used as a *qualitative* measure of sequence similarity.

EFI-EST also provides the global pairwise sequence identity ("%id") as an edge attribute for the SSNs ("alignment score" and "%id" are independent edge attributes). Because of the complications of how BLAST weights gaps, "%id" *is not recommended* for the construction of SSNs for entire protein families (or divergent subsets of clusters from a family SSN) so it is not available for generating the initial SSN using EFI-EST. However, "%id" can be useful for generating isofunctional clusters from heterogeneous SSN clusters that contain "close" homologues. The Select Panel in Cytoscape can be used to select edges based either on their alignment score or "%id" edge attribute; the selected edges can be deleted with the Edit menu and a new SSN generated with the Layout menu.

The strategy for selecting the minimum alignment score for generating the initial SSN is described in the main text; details for implementing this strategy are provided in Section II.E for EFI-EST. A single alignment score threshold may not segregate orthologues into isofunctional clusters across a functionally diverse superfamily because sequence boundaries between functions (substrate specificity and/or reaction mechanism) often do not diverge uniformly as sequence similarity decreases. Orthogonal information, e.g., genome context for microbial and fungal proteins to identify functionally linked proteins in metabolic pathways provided by EFI-GNT, can be used to assess whether clusters are isofunctional. Daughter networks for individual multifunctional clusters in an SSN can be generated with Cytoscape [by selecting/highlighting the cluster nodes and then clicking the "New Network from Selection (all edges)" button at the top of the Cytoscape window]; larger alignment score thresholds then can be applied only to the daughter network to segregate the nodes into multiple clusters that can be assessed for isofunctionality using SwissProt-curated/literature-curated functions and/or genome context.

C. Visualization and Analysis of SSNs Using Cytoscape

The tools provide SSNs in the xgmml file format (uncompressed and zipped); uncompressed files are used as the input for Cytoscape (https://cytoscape.org/), an open source software platform for visualization and analyses of SSNs. The size of the SSN that can be opened is determined by the number of edges and the amount of available RAM on the user's computer; *as an approximate guide*, SSNs with ~2M edges can be opened with 16 GB RAM, ~4 M edges can be opened with 32 GB RAM, ~8M edges can be opened with 64 GB RAM, ~15M edges can be opened with 128 GB RAM, and ~30M edges can be opened with 256 GB RAM. The "**Download Network Files**" page of EFI-EST provides the number of edges in each file so the user can download the highest resolution file that can be opened.

We recommend using the "yFiles Organic Layout" in Cytoscape to visualize SSNs and colored SSNs generated by all three tools. For this reason, we recommend Cytoscape 3.5.1. We find the "yFiles Organic Layout" in this version (and earlier versions) the most informative for visualizing SSNs; in more recent versions, the parameters defining the "yFiles Organic Layout" were changed, so it is more difficult to identify the edges/connections between emerging clusters within SSN clusters. Also, "yFiles Organic Layout" is not installed with more recent versions of Cytoscape, so it must be separately downloaded.

We recommend using the Prefuse Force Directed Layout for both GNNs generated by EFI-GNT. For very large SSNs, the Prefuse Force Directed Layout using the alignment score edge attribute may be preferred over the Organic Layout. It is also possible to use the "None" option in the Prefuse Force Directed Layout; however, the alignment score edge attribute is preferred.

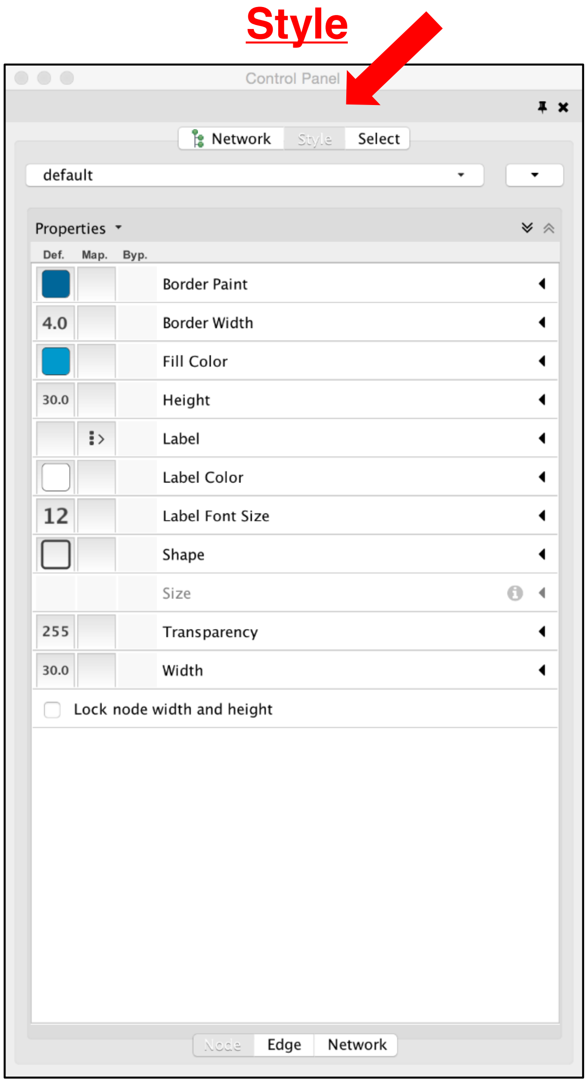
Using the Cytoscape “Select” control panel, edges can be selected based on the value of their “Alignment Score” (or “%id”; *vide supra*) edge attribute and deleted; when subjected to layout, the edited SSNs can be redisplayed to visualize the altered clustering. It often is faster to use the “Dataset Completed” page of EFI-EST (Section II.E of EFI-EST) to generate SSNs at multiple alignment score values if the number of edges is large. We recommend using the “%id” edge attribute for selection/deletion of edges to segregate individual SSN clusters (in daughter networks) that contain sequences with “significant” sequence similarity, e.g., to achieve isofunctionality when paralogues/homologues catalyze the same type of reaction with different substrates.

On the “**Download Network Files**” page (Section II.F of EFI-EST), EFI-EST provides a “full” SSN that contains a node for each UniProt ID (or UniRef cluster ID) as well as “representative node” SSNs (rep node SSNs) in which UniProt IDs (or UniRef cluster IDs) that share from 40 to 100% identity, in 5% intervals, are grouped together in the same metanode. The node attributes for metanodes in rep node SSNs include “Number of IDs in Rep Node” [number of UniProt (or UniRef cluster) IDs] and “List of IDs in Rep Node” (list of the UniProt IDs) as well as lists of the unique values of the node attributes for each UniProt ID in the metanode. If the SSN for the full SSN is too large to be opened with the available RAM, the user should download the highest resolution rep node SSN (greatest sequence identity) that can be opened.

The node attributes provided with SSNs include information about each UniProt ID in the SSN, e.g., “UniProt Annotation Status” (TrEMBL or SwissProt), “Description” (computationally assigned function if TrEMBL), “SwissProt Description” (manually curated function), “Sequence

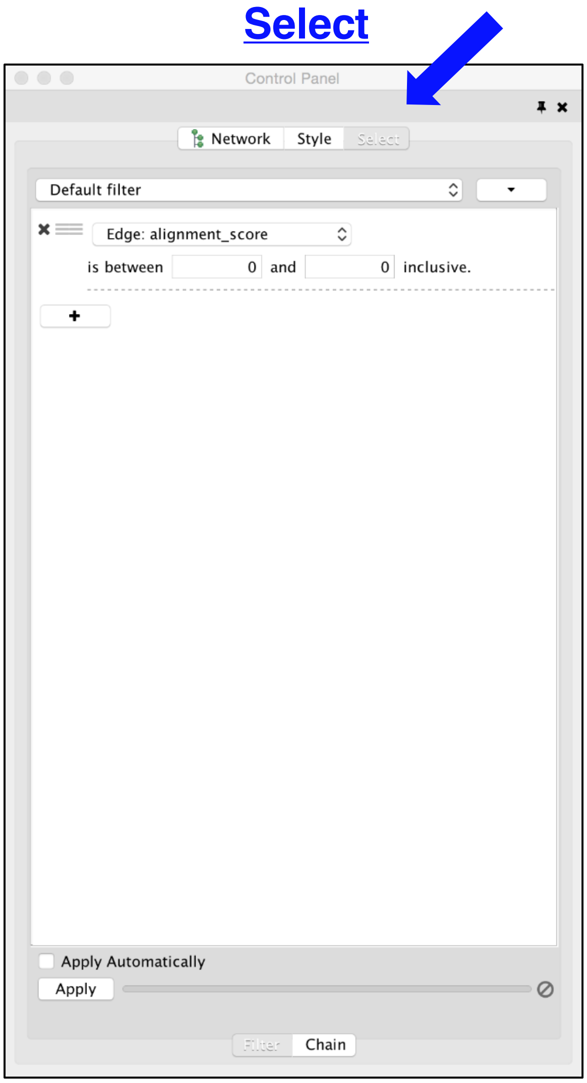
Length", "Organism", phylogenetic classification for the organism ("Superkingdom", "Kingdom", "Phylum", "Class", "Order", "Family", "Genus", and "Species"), and accession codes in databases ("PDB", "PFAM", "InterPro Domain", "InterPro Family", "BRENDA ID", "KEGG ID", "PATRIC ID", "STRING ID", and "GO Term"). The node attributes assist the user in analyzing the SSN and choosing an alignment score for segregating the nodes into isofunctional clusters. The node attributes in the SSNs generated by EFI-EST are listed in Table S1.

Cytoscape provides “Panels” to select and modify nodes to facilitate analyses: 1) “Select” (red arrow) identifies nodes based on the values of their node attributes (text or numerical) and 2) “Style” (blue arrow) changes the color, size, and/or shape of selected nodes.



Def.	Map.	Byp.	Property
🟦			Border Paint
4.0			Border Width
🟦			Fill Color
30.0			Height
			Label
□			Label Color
12			Label Font Size
□			Shape
			Size
255			Transparency
30.0			Width

□ Lock node width and height



Default filter

Edge: alignment_score

is between 0 and 0 inclusive.

□ Apply Automatically

For text node attributes, the types of searches supported by the “Select” panel include “is”, “is not”, “contains”, and “does not contain”; for numerical node attributes, a specific value or a range of values can be selected. Multiple node attributes can be searched; “and” and “or” operators can be applied. Users need to explore both panels—they provide considerable power for both analyzing/editing SSNs and changing the properties of nodes.

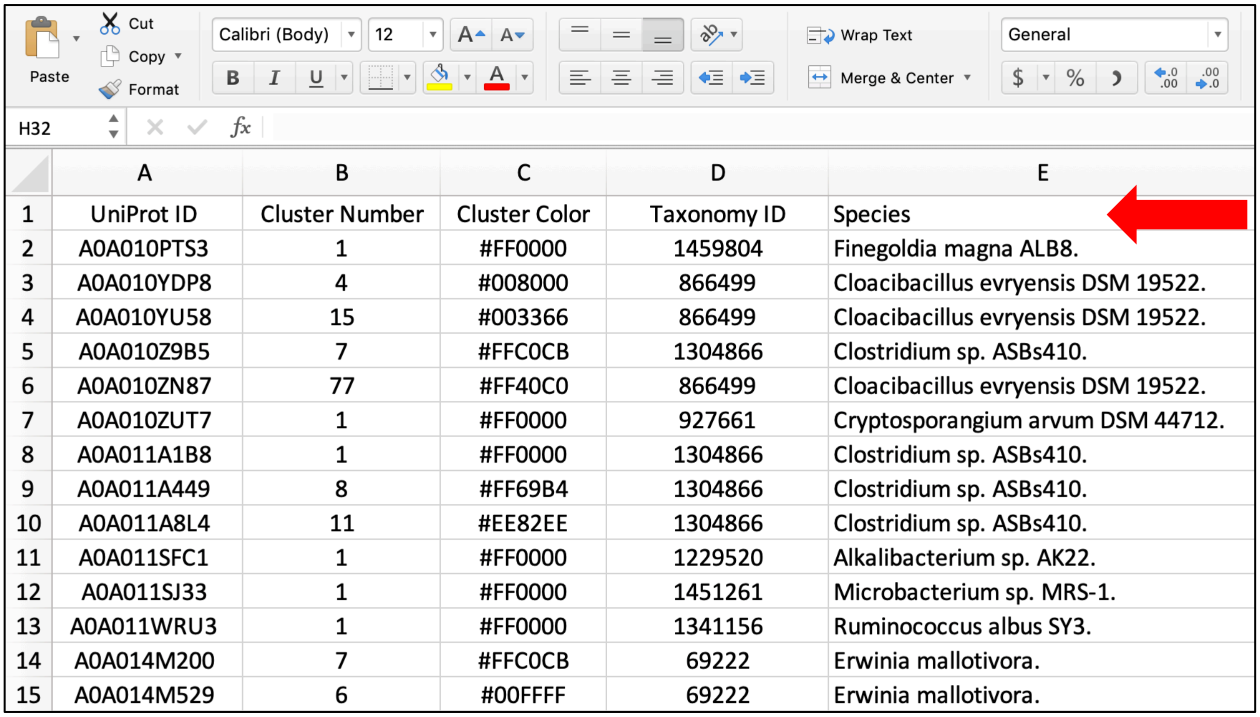
SSNs can be saved as session (cys) files. Daughter networks of specific nodes/clusters in the SSN can be generated after selecting the nodes/clusters in the SSN. Xgmml files for edited and daughter SSNs can be exported for import into the Color SSNs utility of EFI-EST, EFI-GNT, or EFI-CGFP (*vide infra*). The xgmml files can be compressed for quicker uploading.

We provide a brief introduction to using Cytoscape for opening and visualizing SSNs at https://efi.igb.illinois.edu/efi-est/tutorial_cytoscape.php. Users are referred to tutorials provided by Cytoscape (https://github.com/cytoscape/cytoscape-tutorials/wiki) and the Cytoscape User Manual (http://manual.cytoscape.org/en/stable/) for detailed instructions.

D. Cytoscape BridgeDB app: “Custom” Node Attributes

The user may find it useful to add “custom” node attributes to the SSNs generated with EFI-EST. As described in the main text, cross-referencing nodes/clusters in SSNs containing the same IDs but generated with different alignment scores is useful for identifying SSN clusters that may share functional properties, e.g., dehydratases in the glycyl radical enzyme (GRE) superfamily (IPR004184); this can be accomplished by adding custom node attributes that contain the colors and cluster numbers assigned by the Color SSNs utility (Sections II.H and II.I) for a reference SSN (e.g., SSN_{240} in the article). Alternatively, the user may want to add a node attribute for functions reported in the literature but not included in UniProtKB/SwissProt (“SwissProt Description” node attribute in the SSNs generated by EFI-EST).

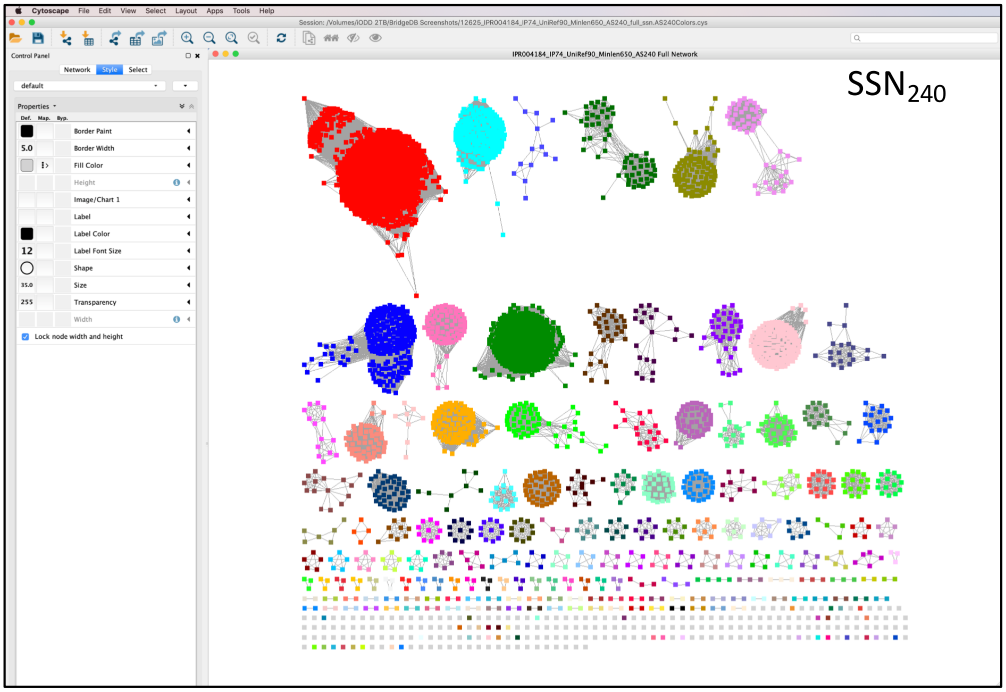
Custom node attributes can be added using the Cytoscape BridgeDB app (http://apps.cytoscape.org/apps/bridgedb). BridgeDB requires a tab-delimited text file with “ID Type” headers in the first row (e.g., red arrow in the example below; from the “UniProt ID-Color-Cluster number” mapping file for SSN240).



	A	B	C	D	E
1	UniProt ID	Cluster Number	Cluster Color	Taxonomy ID	Species
2	A0A010PTS3	1	#FF0000	1459804	Finegoldia magna ALB8.
3	A0A010YDP8	4	#008000	866499	Cloacibacillus evryensis DSM 19522.
4	A0A010YU58	15	#003366	866499	Cloacibacillus evryensis DSM 19522.
5	A0A010Z9B5	7	#FFC0CB	1304866	Clostridium sp. ASBs410.
6	A0A010ZN87	77	#FF40C0	866499	Cloacibacillus evryensis DSM 19522.
7	A0A010ZUT7	1	#FF0000	927661	Cryptosporangium arvum DSM 44712.
8	A0A011A1B8	1	#FF0000	1304866	Clostridium sp. ASBs410.
9	A0A011A449	8	#FF69B4	1304866	Clostridium sp. ASBs410.
10	A0A011A8L4	11	#EE82EE	1304866	Clostridium sp. ASBs410.
11	A0A011SFC1	1	#FF0000	1229520	Alkalibacterium sp. AK22.
12	A0A011SJ33	1	#FF0000	1451261	Microbacterium sp. MRS-1.
13	A0A011WRU3	1	#FF0000	1341156	Ruminococcus albus SY3.
14	A0A014M200	7	#FFC0CB	69222	Erwinia mallotivora.
15	A0A014M529	6	#00FFFF	69222	Erwinia mallotivora.

One “ID Type” (in this example, “UniProt ID”) will associated with a node attribute in the SSN (“Source ID Type”). Typical useful node attributes in the SSN include “name” or “shared name” (UniProt IDs in full networks), “List of IDs in Rep Node” (UniProt IDs in metanodes in rep node SSNs), “UniRef50 Cluster IDs” or “UniRef90 Cluster IDs” [SSNs generated using UniRef90 (this example; *vide infra*) and UniRef50 clusters], “Taxonomy ID” (NCBI taxonomy IDs for organisms), and “Species”. The other “ID Type(s)” in the tab-delimited text file specify the custom node attribute(s).

The “UniProt ID-Color-Cluster number” mapping tab-delimited text file was generated by the Color SSNs utility using SSN_{240}; the file associates each UniProt ID in SSN_{240} with the unique SSN cluster color and number assigned by the utility. In this example, the node/cluster colors in the colored SSN for SSN_{240} will be associated with nodes in SSN_{185}.

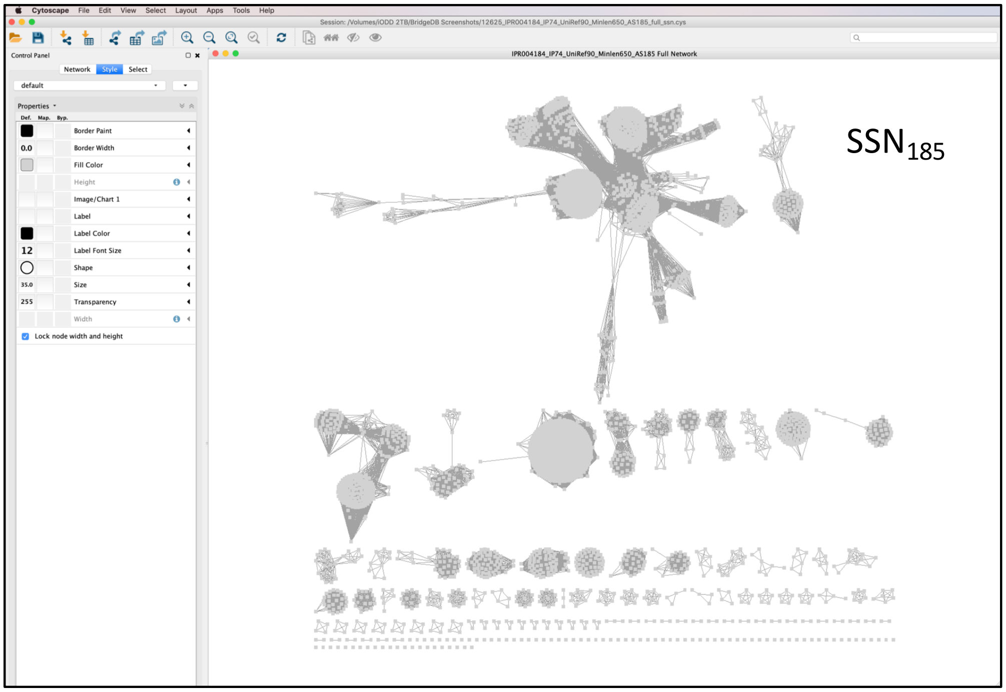


Session: /Volumes/iODD 2TB/BridgeDB Screenshots/12625_IPR004184_IP74_UniRef90_Minlen650_AS240_full_ssn.AS240Colors.cys

IPR004184_IP74_UniRef90_Minlen650_AS240 Full Network

Def.	Map.	Byp.	
■			Border Paint
5.0			Border Width
□	⁝>		Fill Color
			Height
			Image/Chart 1
			Label
■			Label Color
12			Label Font Size
○			Shape
35.0			Size
255			Transparency
			Width

☑ Lock node width and height



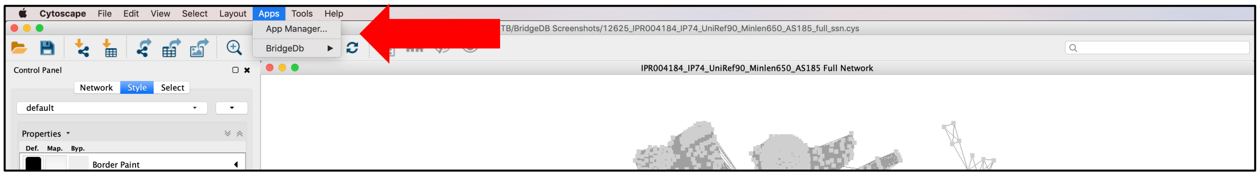
Session: /Volumes/iODD 2TB/BridgeDB Screenshots/12625_IPR004184_IP74_UniRef90_Minlen650_AS185_full_ssn.cys

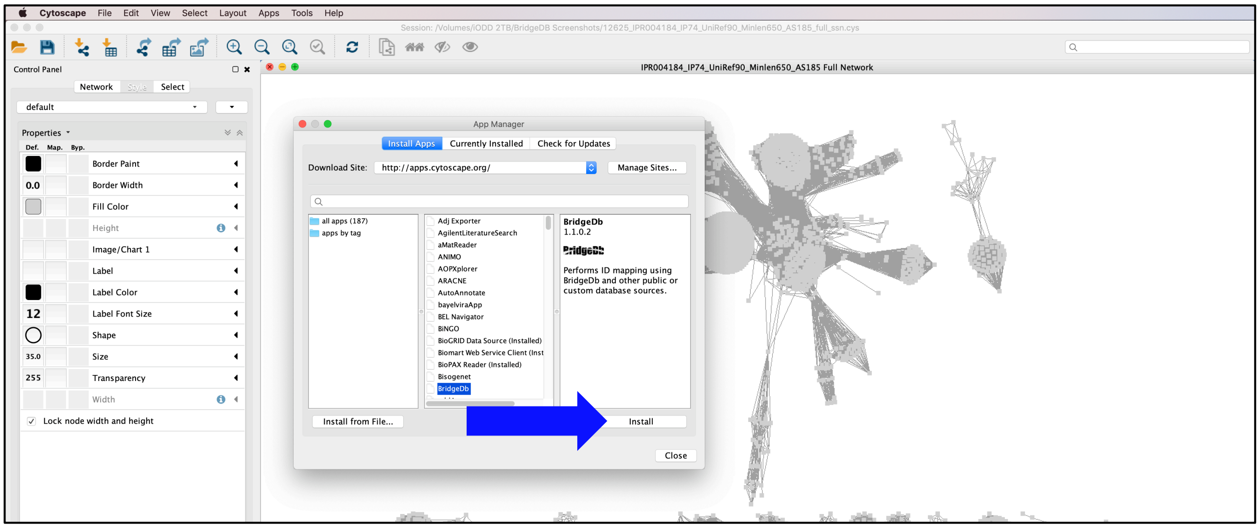
IPR004184_IP74_UniRef90_Minlen650_AS185 Full Network

Def.	Map.	Byp.	
■			Border Paint
0.0			Border Width
□			Fill Color
			Height
			Image/Chart 1
			Label
■			Label Color
12			Label Font Size
○			Shape
35.0			Size
255			Transparency
			Width

☑ Lock node width and height

The Cytoscape “Apps” menu contains “App Manager” (red arrow). Selecting it opens the App Manager window. BridgeDB is selected and then installed by clicking “Install” (blue arrow).

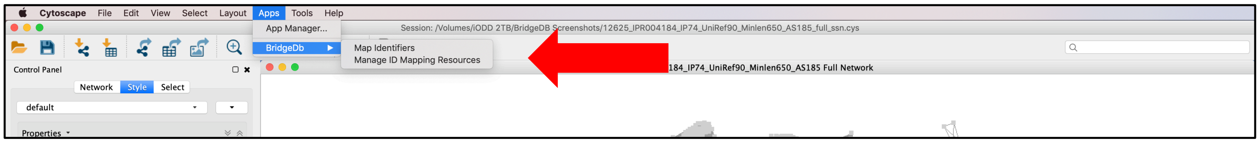




BridgeDb
1.1.0.2

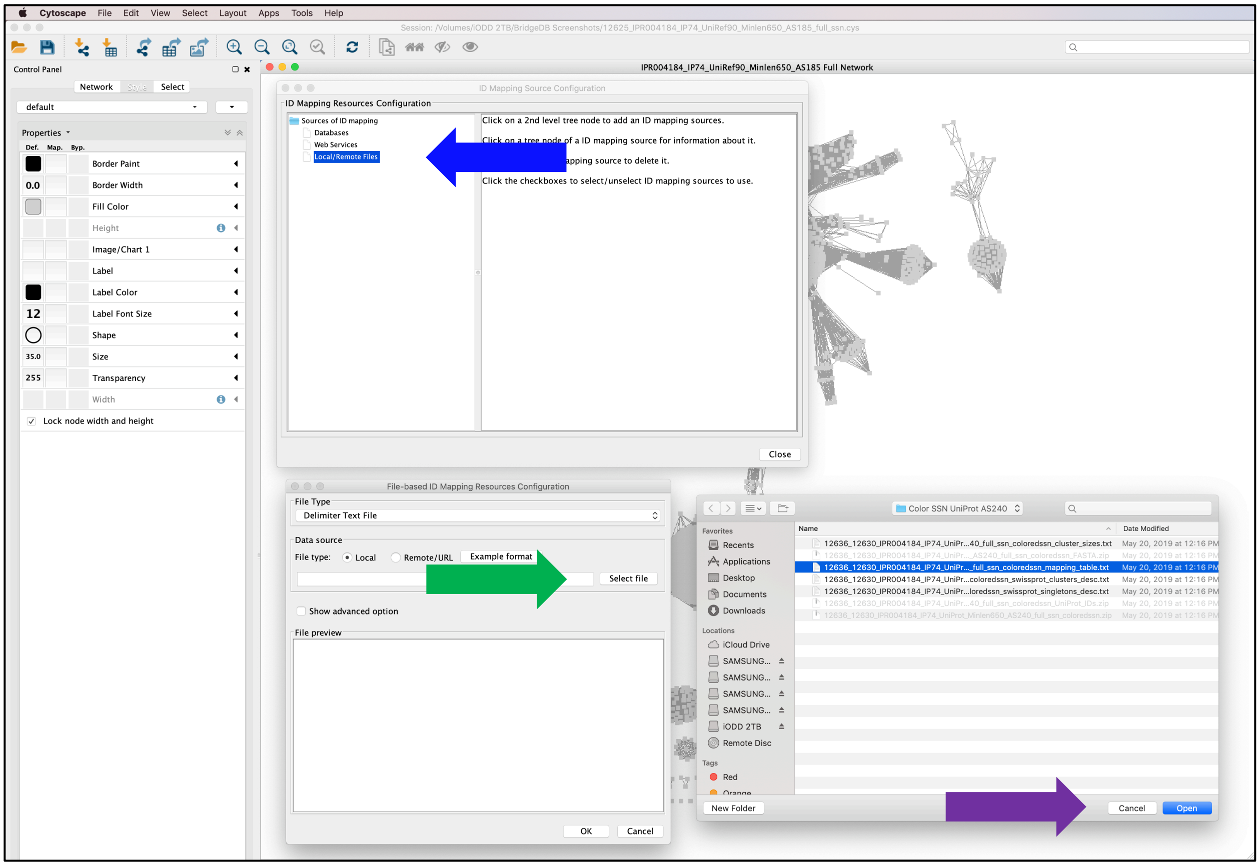
Performs ID mapping using BridgeDb and other public or custom database sources.

The "Apps" menu now contains "BridgeDB" with two submenus: "Map Identifiers" and "Manage ID Mapping Resources" (red arrow). By selecting "Manage ID Mapping Resources", the "ID Mapping Resources Configuration" window opens. Selecting "Local/Remote Files" (blue arrow) opens the window that allows the mapping file to be selected (green arrow). After the file is located/selected, it is uploaded by clicking "Open" (magenta arrow).



Session: /Volumes/iODD 2TB/BridgeDB Screenshots/12625_IPR004184_IP74_UniRef90_Minlen650_AS185_full_ssn.cys

184_IP74_UniRef90_Minlen650_AS185 Full Network



Session: /Volumes/iODD 2TB/BridgeDB Screenshots/12625_IPR004184_IP74_UniRef90_Minlen650_AS185_full_ssn.cys

IPR004184_IP74_UniRef90_Minlen650_AS185 Full Network

ID Mapping Source Configuration

ID Mapping Resources Configuration

- Sources of ID mapping
  - Databases
  - Web Services
  - Local/Remote Files

Click on a 2nd level tree node to add an ID mapping sources.

Click on a tree node of a ID mapping source for information about it.

apping source to delete it.

Click the checkboxes to select/unselect ID mapping sources to use.

File-based ID Mapping Resources Configuration

File Type: Delimiter Text File

Data source

File type: ● Local ○ Remote/URL

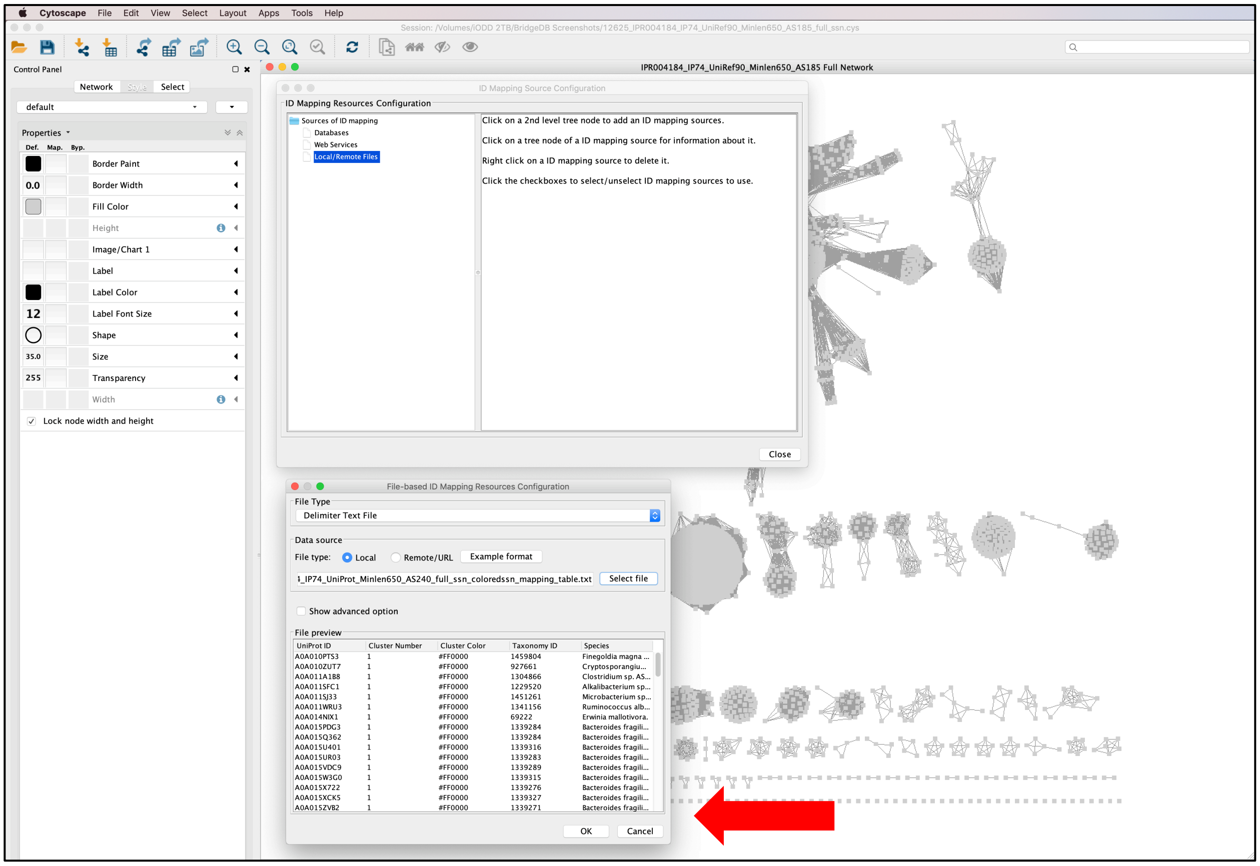
□ Show advanced option

File preview

Color SSN UniProt AS240

Name	Date Modified
12636_12630_IPR004184_IP74_UniPr...40_full_ssn_coloredssn_cluster_sizes.txt	May 20, 2019 at 12:16 PM
12636_12630_IPR004184_IP74_UniPr..._AS240_full_ssn_coloredssn_FASTA.zip	May 20, 2019 at 12:16 PM
12636_12630_IPR004184_IP74_UniPr..._full_ssn_coloredssn_mapping_table.txt	May 20, 2019 at 12:16 PM
12636_12630_IPR004184_IP74_UniPr...coloredssn_swissprot_clusters_desc.txt	May 20, 2019 at 12:16 PM
12636_12630_IPR004184_IP74_UniPr...loredssn_swissprot_singletons_desc.txt	May 20, 2019 at 12:16 PM
12636_12630_IPR004184_IP74_UniPr...40_full_ssn_coloredssn_UniProt_IDs.zip	May 20, 2019 at 12:16 PM
12636_12630_IPR004184_IP74_UniProt_Minlen650_AS240_full_ssn_coloredssn.zip	May 20, 2019 at 12:16 PM

The "File Type" window then provides a preview of the uploaded file. If correct, click "OK". The "Mapping Resources Configuration" window opens with the uploaded file and a check box that shows that the file is selected (blue arrow). Multiple files can be uploaded and saved; the desired file is selected. The window is closed by clicking "Close" (green arrow).


Cytoscape File Edit View Select Layout Apps Tools Help
Session: /Volumes/iODD 2TB/BridgeDB Screenshots/12625_IPR004184_IP74_UniRef90_Minlen650_AS185_full_ssn.cys
Control Panel
Network Style Select
default
Properties
Def. Map. Byp.
Border Paint
0.0 Border Width
Fill Color
Height
Image/Chart 1
Label
Label Color
12 Label Font Size
Shape
35.0 Size
255 Transparency
Width
Lock node width and height
IPR004184_IP74_UniRef90_Minlen650_AS185 Full Network

ID Mapping Source Configuration

ID Mapping Resources Configuration

- Sources of ID mapping
  - Databases
  - Web Services
  - Local/Remote Files

Click on a 2nd level tree node to add an ID mapping sources.

Click on a tree node of a ID mapping source for information about it.

Right click on a ID mapping source to delete it.

Click the checkboxes to select/unselect ID mapping sources to use.

Close

File-based ID Mapping Resources Configuration

File Type: Delimiter Text File

Data source

File type: ● Local ○ Remote/URL Example format

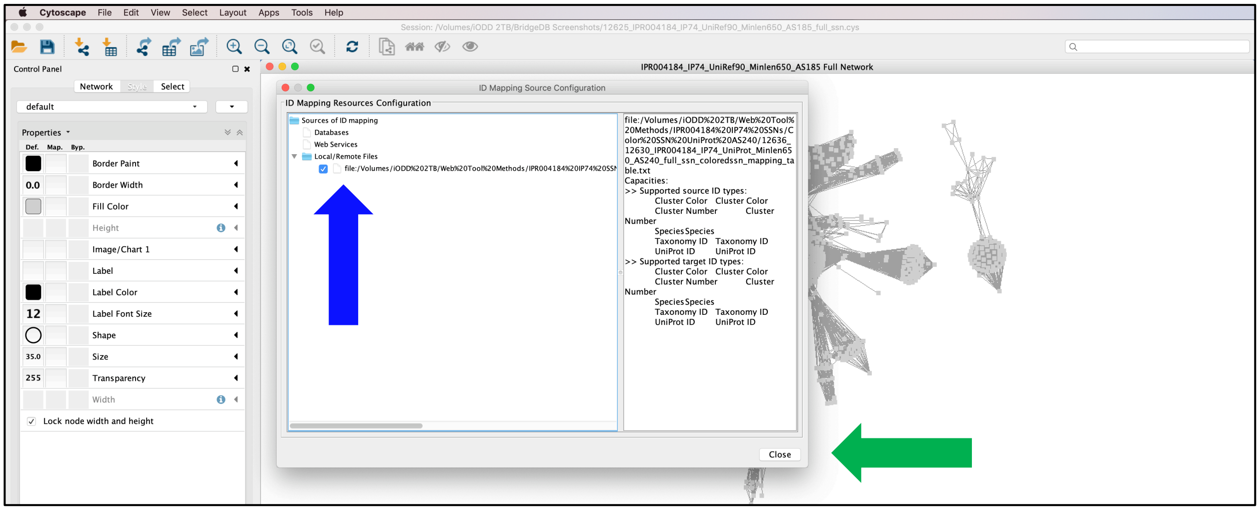
1_IP74_UniProt_Minlen650_AS240_full_ssn_coloredssn_mapping_table.txt Select file

□ Show advanced option

File preview

UniProt ID	Cluster Number	Cluster Color	Taxonomy ID	Species
A0A010PTS3	1	#FF0000	1459804	Finegoldia magna ...
A0A010ZUT7	1	#FF0000	927661	Cryptosporangiu...
A0A011A1B8	1	#FF0000	1304866	Clostridium sp. AS...
A0A011SFC1	1	#FF0000	1229520	Alkalibacterium sp...
A0A011SJ33	1	#FF0000	1451261	Microbacterium sp...
A0A011WRU3	1	#FF0000	1341156	Ruminococcus alb...
A0A014NIX1	1	#FF0000	69222	Erwinia mallotivora.
A0A015PDG3	1	#FF0000	1339284	Bacteroides fragili...
A0A015Q362	1	#FF0000	1339284	Bacteroides fragili...
A0A015U401	1	#FF0000	1339316	Bacteroides fragili...
A0A015UR03	1	#FF0000	1339283	Bacteroides fragili...
A0A015VDC9	1	#FF0000	1339289	Bacteroides fragili...
A0A015W3G0	1	#FF0000	1339315	Bacteroides fragili...
A0A015X722	1	#FF0000	1339276	Bacteroides fragili...
A0A015XCK5	1	#FF0000	1339327	Bacteroides fragili...
A0A015ZVB2	1	#FF0000	1339271	Bacteroides fragili...

OK Cancel


Cytoscape File Edit View Select Layout Apps Tools Help
Session: /Volumes/iODD 2TB/BridgeDB Screenshots/12625_IPR004184_IP74_UniRef90_Minlen650_AS185_full_ssn.cys
Control Panel
Network Style Select
default
Properties
Def. Map. Byp.
Border Paint
0.0 Border Width
Fill Color
Height
Image/Chart 1
Label
Label Color
12 Label Font Size
Shape
35.0 Size
255 Transparency
Width
Lock node width and height
IPR004184_IP74_UniRef90_Minlen650_AS185 Full Network

ID Mapping Source Configuration

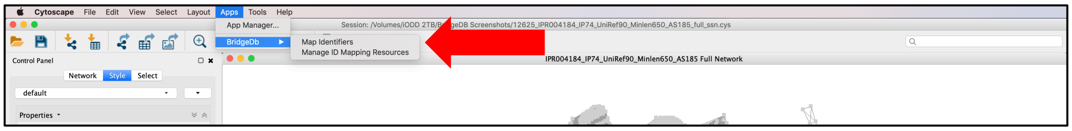
ID Mapping Resources Configuration

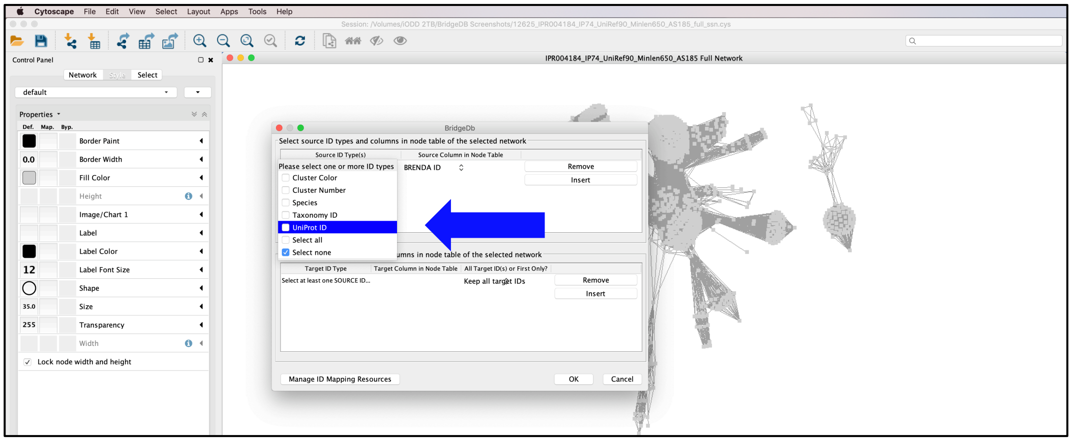
- Sources of ID mapping
  - Databases
  - Web Services
  - Local/Remote Files
    - ☑ file:/Volumes/iODD%202TB/Web%20Tool%20Methods/IPR004184%20IP74%20SSN

```
file:/Volumes/iODD%202TB/Web%20Tool%
20Methods/IPR004184%20IP74%20SSNs/C
olor%20SSN%20UniProt%20AS240/12636_
12630_IPR004184_IP74_UniProt_Minlen65
0_AS240_full_ssn_coloredssn_mapping_ta
ble.txt
Capacities:
>> Supported source ID types:
        Cluster Color   Cluster Color
        Cluster Number          Cluster
Number
        Species Species
        Taxonomy ID   Taxonomy ID
        UniProt ID        UniProt ID
>> Supported target ID types:
        Cluster Color   Cluster Color
        Cluster Number          Cluster
Number
        Species Species
        Taxonomy ID   Taxonomy ID
        UniProt ID        UniProt ID
```

Close

The custom node attributes are then added to the SSN. “BridgeDB” in the “Apps” menu includes “Map Identifiers” (red arrow). Selecting this submenu opens a window.


Cytoscape File Edit View Select Layout Apps Tools Help
App Manager...
BridgeDb
Map Identifiers
Manage ID Mapping Resources
Session: /Volumes/iODD 2TB/BridgeDB Screenshots/12625_IPR004184_IP74_UniRef90_Minlen650_AS185_full_ssn.cys
IPR004184_IP74_UniRef90_Minlen650_AS185 Full Network
Control Panel
Network Style Select
default
Properties

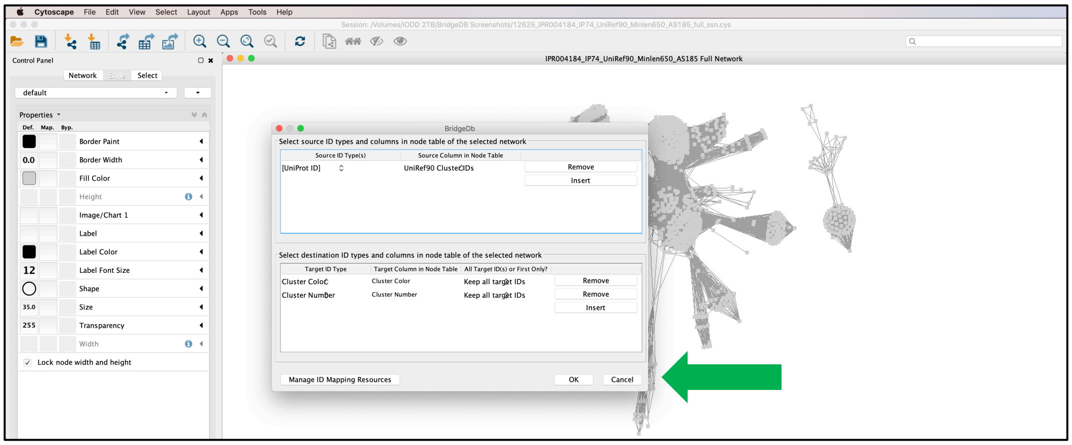


Select source ID types and columns in node table of the selected network

Source ID Type(s)	Source Column in Node Table
Please select one or more ID types	BRENDA ID

- □ Cluster Color
- □ Cluster Number
- □ Species
- □ Taxonomy ID
- □ UniProt ID
- □ Select all
- ☑ Select none

Target ID Type	Target Column in Node Table	All Target ID(s) or First Only?
Select at least one SOURCE ID...		Keep all target IDs

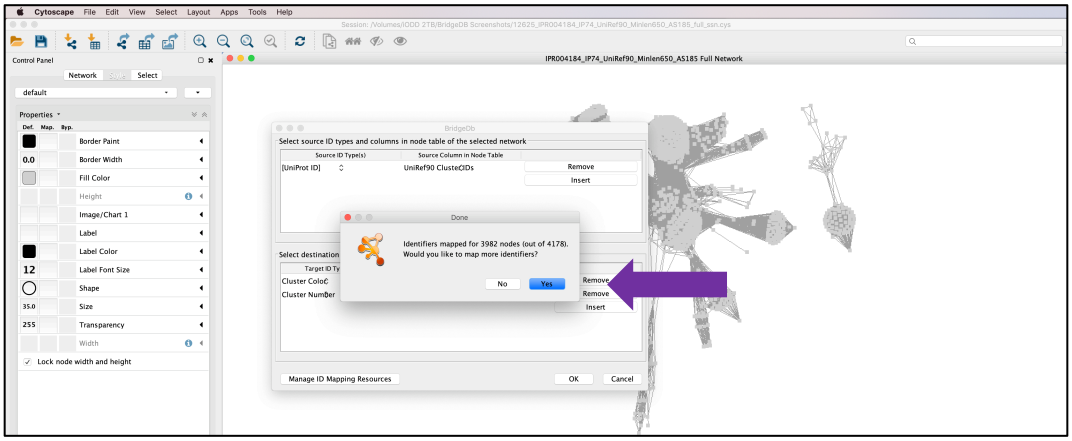


Select source ID types and columns in node table of the selected network

Source ID Type(s)	Source Column in Node Table
[UniProt ID]	UniRef90 ClusterIDs

Select destination ID types and columns in node table of the selected network

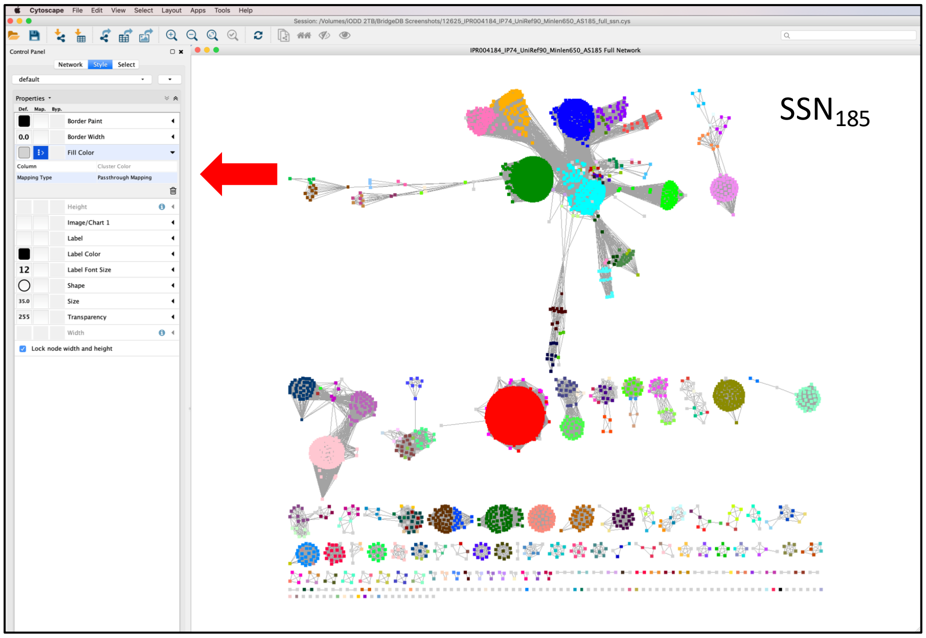
Target ID Type	Target Column in Node Table	All Target ID(s) or First Only?
Cluster Color	Cluster Color	Keep all target IDs
Cluster Number	Cluster Number	Keep all target IDs



Identifiers mapped for 3982 nodes (out of 4178).
Would you like to map more identifiers?

Selecting the “Please select one or more ID types” entry under “Source ID Type(s)” header opens a panel with a list of the ID types (column headers) in the uploaded mapping file (blue arrow). Select the ID type (“UniProt ID”) that will be matched with a node attribute in SSN_{185}; under the “Source Column in Node Attribute” header select the SSN_{185} “UniRef90 Cluster IDs” node attribute that completes the match. In the lower panel of the window, select the “Source ID types” in the mapping file that will be added as the custom node attributes (“Cluster Color” and “Cluster Number”); clicking on the “Insert” button allows multiple “Source ID Types” to be added. When finished, click the “OK” button (green arrow). BridgeDB then matches the values of “UniProt ID” in the mapping file with nodes that contain the same values in the “UniRef90 Cluster IDs” node attribute [SSN_{185} was generated using UniRef90 clusters IDs (*vide infra*); the “UniRef90 Cluster IDs” node attribute contains the UniProt IDs in each UniRef90 cluster]. A window opens when the matching is completed; click “OK” to finish (magenta arrow). The Table Panel in the Cytoscape window now contains the “Cluster Color” and “Cluster Number” node attributes.

The Cytoscape "Style Panel" includes the "Fill Color" node property. Click on the triangle; then select the "Cluster Color" node attribute in the "Column" line; select "Passthrough Mapping" in the "Mapping Type" line (red arrow). After a few seconds, the nodes in SSN$_{185}$ are colored according to the colors specified by the "Cluster Color" node attribute (from the colored SSN$_{240}$).



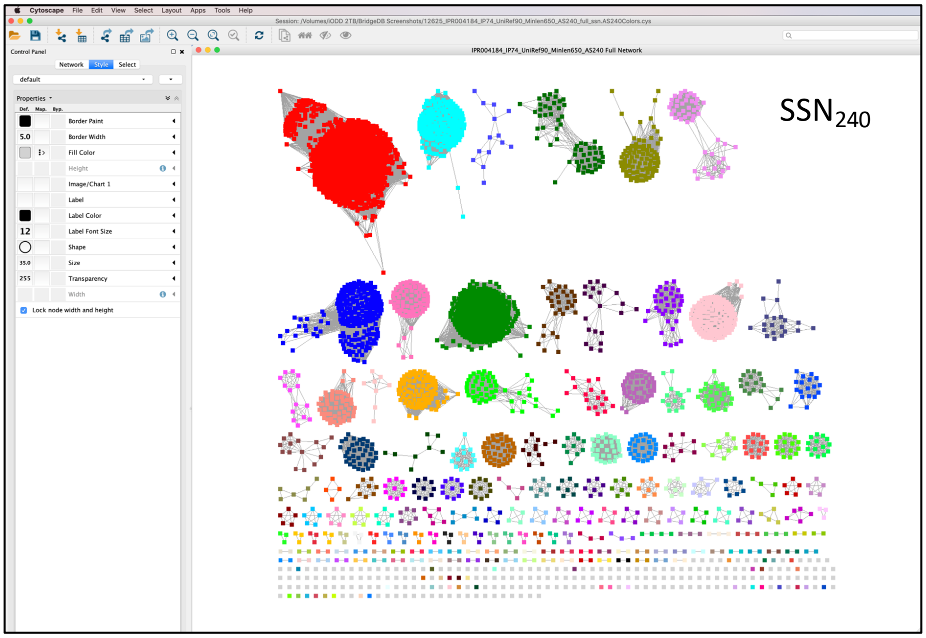
Session: /Volumes/iODD 2TB/BridgeDB Screenshots/12625_IPR004184_IP74_UniRef90_Minlen650_AS185_full_ssn.cys

IPR004184_IP74_UniRef90_Minlen650_AS185 Full Network

Fill Color

Column: Cluster Color

Mapping Type: Passthrough Mapping



Session: /Volumes/iODD 2TB/BridgeDB Screenshots/12625_IPR004184_IP74_UniRef90_Minlen650_AS240_full_ssn.AS240Colors.cys

IPR004184_IP74_UniRef90_Minlen650_AS240 Full Network

E. SSNs for Large Protein Families: UniProt and UniRef Databases

As the protein databases continue to grow (**Figure S1** in the Supplementary Figures), the tools must be compatible with the increasing sizes of protein families—the computational requirements for calculation of edges increase, and the ability to visualize the SSNs becomes more challenging. The former can be addressed with an increase in the number of processors; the latter with an increase in the amount of computer system RAM for Cytoscape.

A solution to both problems is to cluster multiple UniProt IDs into metanodes based on sequence similarity, with edge calculations and visualization both using the sequence of the metanode identifier instead of the multiple sequences within the metanode. In addition to the UniProt/TrEMBL and UniProt/SwissProt databases with individual sequences, UniProt provides the UniProt Reference Cluster (UniRef) database in which UniProt entries are grouped in clusters (this use of "cluster" can be confusing in the context of SSNs because "cluster" is also used to describe nodes in SSNs that are connected by edges) that share 100%, 90% or 50% sequence identity (https://www.uniprot.org/help/uniref). The UniRef100 database (not used by EFI-EST) combines identical sequences and fragments with ≥11 residues into a single UniRef90 cluster (about 80% of the sequences in UniProt are unique). The UniRef90 database (used by EFI-EST) groups UniRef100 cluster seed sequences (longest sequence in the cluster, not necessarily the "most informative" sequence that is used as the UniRef cluster ID; *vide infra*) that share ≥90% sequence identity and ≥80% overlap with the longest (seed) sequence in the UniRef90 cluster; on average, the number of UniRef90 clusters in a Pfam family is ~50% the number of UniProt IDs. The UniRef50 database (used by EFI-EST) clusters UniRef90 cluster seed sequences (longest sequence, not necessarily the "most informative" sequence that is used as the UniRef cluster ID; *vide infra*) that share ≥50% sequence identity and ≥80% overlap with the longest (seed) sequence in the UniRef50 cluster; on average, the number of UniRef50 clusters in a Pfam family is ~20% the number of UniProt IDs. The sequences in a UniRef90 cluster likely are orthologues; many sequences in a UniRef50 cluster are orthologues, but some may be functionally distinct homologues.

The SSNs generated by EFI-EST provide the UniProt IDs in each UniRef90 cluster and the UniProt IDs and UniRef90 cluster IDs in each UniRef50 cluster. The SSNs generated with UniRef90 and UniRef50 clusters include "UniRef90 Cluster IDs" and "UniRef50 Cluster IDs" node attributes that provide lists of the UniProt IDs in each metanode [UniProt IDs grouped by sequence identity in multiply populated UniRef clusters and/or metanodes in representative node (rep node) SSNs; *vide infra*]; they also contain the "Cluster ID Sequence Length" node attribute. The other SSN node attributes generated with the UniRef90 and UniRef50 databases are lists of the values for all of the UniProt IDs in the UniRef cluster/SSN metanodes, including the "Sequence Length" node attribute with a list of the lengths of all UniProt IDs in the metanode. The Color SSNs utility (Section II.H for EFI-EST) provides folders of files for each SSN cluster that includes 1) the UniProt IDs for each SSN cluster in an SSN generated with the UniRef90 database and 2) the UniProt IDs and UniRef90 cluster IDs for each SSN cluster in an SSN generated with the UniRef50 database; these can be as the input for Option D of EFI-EST to generate higher resolution SSNs for clusters of interest in UniRef90 and UniRef50 SSNs.

The use of UniRef clusters to generate SSNs reduces the number of sequences used in the BLAST and, therefore, the time required for the pairwise comparisons; relative to using UniProt sequences, on average the time required for the BLAST is reduced by ~75% for UniRef90 clusters and ~96% for UniRef50 clusters. The use of UniRef clusters also reduces the number of SSN nodes (metanodes), thereby allowing larger families to be visualized with laptop/desktop computers with commonly available amounts of RAM (≤16 GB RAM). The size of the SSN that can be opened is determined by the number of edges (and, therefore, the number of nodes that are connected by edges) and the amount of available RAM on the user's computer. *As an approximate guide*, SSNs with ~2M edges can be opened with 16 GB RAM, ~4 M edges can be opened with 32 GB RAM, ~8M edges can be opened with 64 GB RAM, ~15M edges can be opened with 128 GB RAM, and ~30M edges can be opened with 256 GB RAM.

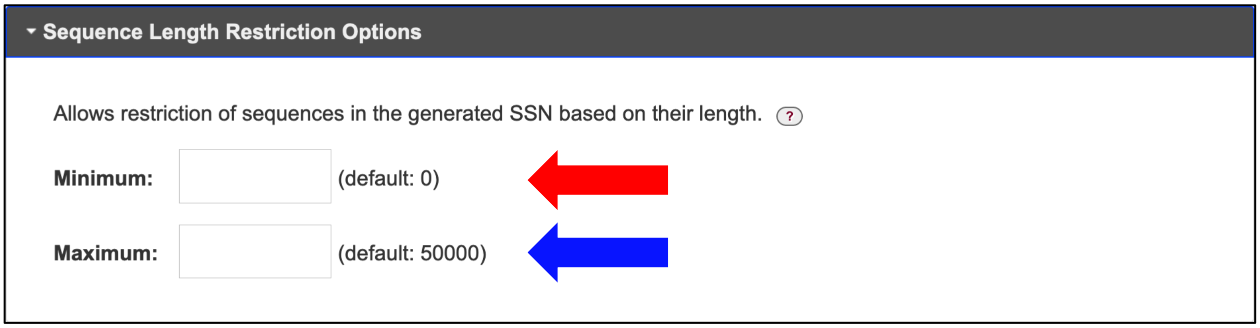
When the UniProt database (UniProt IDs) is used to generate SSNs, EFI-EST identifies and uses unique sequences for the BLAST; the values of these edges are applied to the redundant sequences so that the SSN includes all sequences/edges in the input dataset. When UniRef90 or UniRef50 clusters are used, the "most informative" sequence in each cluster (as defined by UniRef; the cluster ID) is used for the BLAST; the criteria for "most informative" in order of importance are 1) quality (SwissProt is preferred); 2) UniProt annotation score (measure of annotation content); 3) organism (reference proteomes are preferred), and 4) length (longest is preferred).

For families with >25,000 UniProt IDs, EFI-EST *requires* the use of UniRef90 clusters to conserve computational resources and enhance the likelihood that the full SSN (all sequences) can be opened and visualized. In UniProt Release 2019_04, 1,666 of the 17,929 Pfam families include

>25,000 UniProt IDs. The node attributes in UniRef90 SSNs include “UniRef90 Cluster Size” and “UniRef90 Cluster IDs”, the latter allowing the user to find any UniProt ID represented in the SSN using the Cytoscape Select panel. The node attributes in UniRef90 SSNs also include “UniRef90 Cluster ID Sequence Length”. As in rep node SSNs (*vide infra*), the other node attributes are lists of the unique values for the individual UniProt IDs in the UniRef90 cluster. For Pfam families with >100,000 UniRef90 clusters (68 families in UniProt Release 2019_04), EFI-EST *requires* the use of UniRef50 clusters; a maximum of 100,000 UniRef50 clusters is imposed. The node attributes in UniRef50 SSNs include “UniRef50 Cluster Size” and “UniRef50 Cluster IDs”, the latter allowing the user to find any UniProt ID represented in the SSN using the Cytoscape Select panel. The node attributes in UniRef50 SSNs also include “UniRef90 Cluster ID Sequence Length”. Again, the other node attributes are lists of the unique values for the individual UniProt IDs in the UniRef50 cluster. Six Pfam families in UniProt Release 2019_04 include >100,000 UniRef50 clusters [PF02518 (HATPase_c), PF00072 (Response_reg), PF00069 (Pkinase), PF00005 (ABC_Tran), PF00512 (HisKA), and PF07690 (MFS_1)]; users interested in generating SSNs for these Pfam families should contact the authors.

UniRef90 and UniRef50 clusters contain fragments [because UniRef100 clusters include short sequences (≥11 residues) that share 100% identity with the seed sequence; *vide infra*]; in UniRef90 and UniRef50 SSNs, the “Sequence Length” node attribute provides a list of the length of all UniProt IDs in the cluster. Fragments have no impact on SSNs—their presence does not influence the values of the edges (these are determined using the UniRef cluster IDs) but will add to the values for the node attributes (lists of the unique values for all UniProt IDs in a metanode). Fragments should have minimal impact on the GNNs generated by EFI-GNT—they often are located at the ends of contigs (truncated sequences), so genome context will be available in only one direction from the query sequence; in large-scale analyses, the co-occurrence frequencies of queries and genome neighbors should not be significantly affected. However, fragments will bias the consensus sequences generated by EFI-CGFP in the identification of the unique ShortBRED family sequence motifs (markers) that are used to quantify metagenome abundance (*vide infra*). Fragments can be removed by the **"Sequence Length Restrictions Options"** when SSNs are generated (next paragraph) and/or when SSNs are uploaded to EFI-CGFP (two paragraphs below).

The "**Sequence Length Restrictions Options**" on the "**SSN Finalization**" tab on "**Dataset Complete**d" page (Section II.E for EFI-EST) provides the ability to select sequences based on user-provided minimum and/or maximum length values to exclude fragments or multidomain proteins or include only specific domain architectures.


▾ Sequence Length Restriction Options

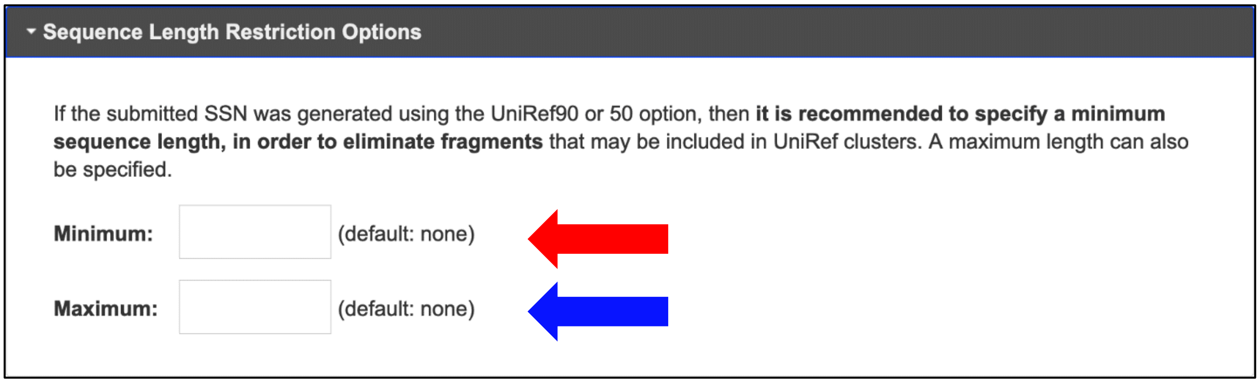
Allows restriction of sequences in the generated SSN based on their length. ?

Minimum: (default: 0)

Maximum: (default: 50000)

For SSNs generated with UniRef90 or UniRef50 cluster IDs, the filters are applied to both the sequences in the clusters as well as the cluster IDs. First, the sequences in each UniRef cluster are length-filtered; those that do not satisfy the requirements are removed from the cluster. Then, the UniRef cluster ID sequences are length-filtered; those that do not satisfy the criteria are discarded. A UniRef cluster may contain sequences longer than the cluster ID (e.g., the "most informative" sequence in the cluster, not necessarily the longest sequence) that exceed the minimum length filter even if the cluster ID sequence does not; however, because UniRef clusters are discarded based the length of their cluster IDs, these "acceptable" sequences are discarded.

As noted above, "**Sequence Length Restriction Options**" also are provided on the "**Run CGFP/ShortBRED**" tab on the EFI-CGFP home page (Section IV.A for EFI-CGFP) so that the user can ensure that the consensus sequences for the ShortBRED families used to identity unique sequence markers are not biased by the presence of fragments.

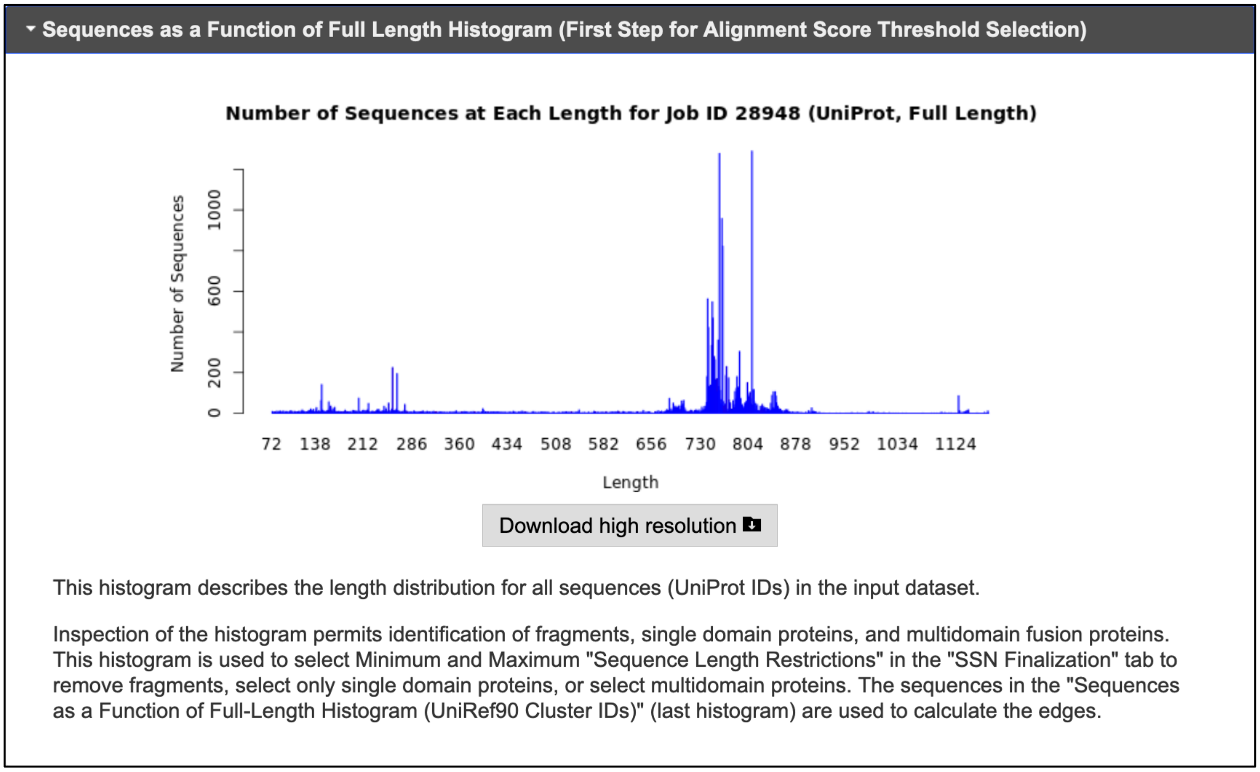


If the submitted SSN was generated using the UniRef90 or 50 option, then **it is recommended to specify a minimum sequence length, in order to eliminate fragments** that may be included in UniRef clusters. A maximum length can also be specified.

Minimum: (default: none)

Maximum: (default: none)

The "**Dataset Analysis**" tab on the "**Dataset Completed**" page (Section II.E.1 for EFI-EST) provides the "**Sequences as a Function of Full-Length Histogram (First Step for Alignment Score Threshold Selection)**" for the full set of UniProt IDs (or sequences in the FASTA files used as input for Option C) in the input dataset, irrespective of whether the input dataset used UniProt IDs, UniProt90 cluster IDs, or UniProt50 cluster IDs. This histogram should be used to evaluate the minimum length of full-length single domain proteins for selecting the minimum alignment score threshold used to generate the initial SSN as well as the presence of fragments and multidomain architectures as deduced from sequence lengths.

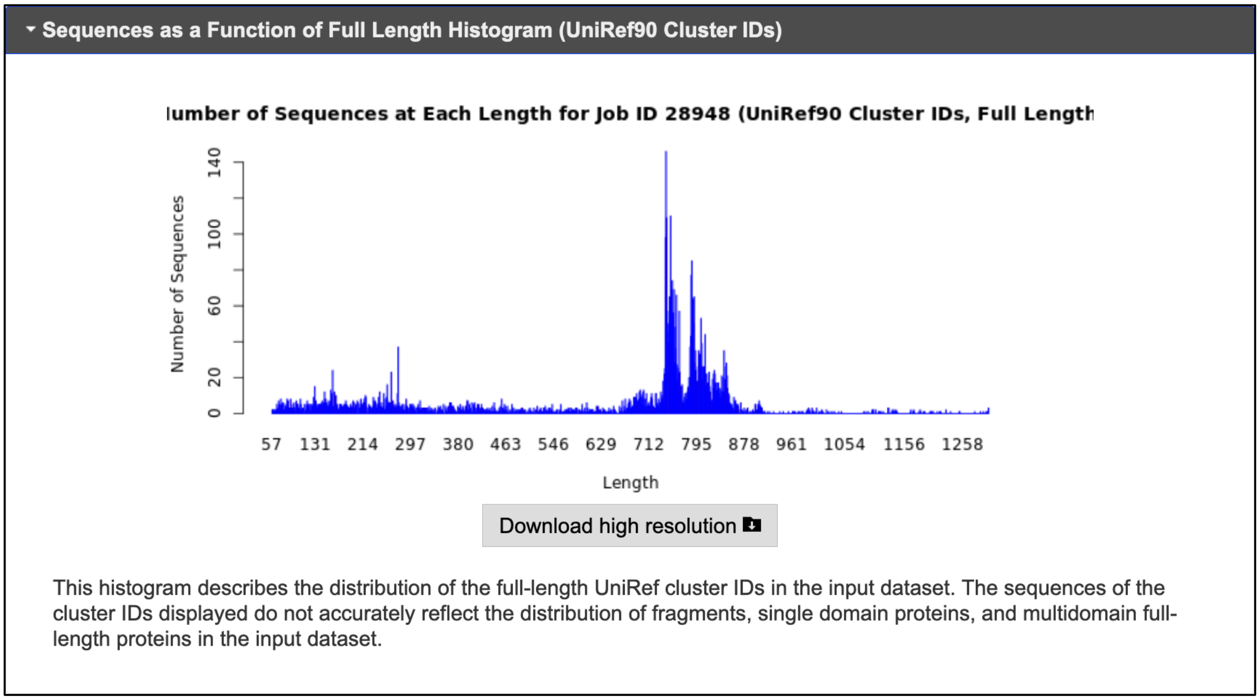


▾ Sequences as a Function of Full Length Histogram (First Step for Alignment Score Threshold Selection)

This histogram describes the length distribution for all sequences (UniProt IDs) in the input dataset.

Inspection of the histogram permits identification of fragments, single domain proteins, and multidomain fusion proteins. This histogram is used to select Minimum and Maximum "Sequence Length Restrictions" in the "SSN Finalization" tab to remove fragments, select only single domain proteins, or select multidomain proteins. The sequences in the "Sequences as a Function of Full-Length Histogram (UniRef90 Cluster IDs)" (last histogram) are used to calculate the edges.

The “**Dataset Analysis**” tab on the “**Dataset Completed**” page (Section II.E.1 for EFI-EST) also provides the “**Sequences as a Function of Full-Length Histogram (First Step for Alignment Score Threshold Selection)**” for the UniRef cluster IDs when UniRef90 or UniRef50 clusters are used to generate the SSN. This histogram overestimates the contribution by fragments (because these have diverse sequences and lengths), although it contains the sequences used in the pairwise comparisons that generate the alignment scores.


▾ Sequences as a Function of Full Length Histogram (UniRef90 Cluster IDs)
lumber of Sequences at Each Length for Job ID 28948 (UniRef90 Cluster IDs, Full Length
Number of Sequences
0
20
60
100
140
57
131
214
297
380
463
546
629
712
795
878
961
1054
1156
1258
Length
Download high resolution

This histogram describes the distribution of the full-length UniRef cluster IDs in the input dataset. The sequences of the cluster IDs displayed do not accurately reflect the distribution of fragments, single domain proteins, and multidomain full-length proteins in the input dataset.

F. SSNs for Multidomain Proteins

Proteins often contain multiple domains, e.g., transcriptional regulators with ligand and DNA-binding domains, radical SAM enzymes with the domain to generate the 5'-deoxyadenosine radical and accessory domains, e.g., the SPASM and B12-binding domains, to enable complex reaction mechanisms, and nonribosomal peptide synthases (NRPSs) that assemble amino acids into complex structures using a variety of catalytic domains (including adenylation, condensation, epimerase, reductase, and thioesterase). By default, EFI-EST generates SSNs for full-length multidomain proteins, with the edge values based on alignment of the full-length sequences, not the domain that is specified by the input family. However, if the input family is heterogeneous with respect to domain architecture, interpretation of the SSN may be difficult because the alignment scores include contributions from, and may be dominated by, the "ancillary" domains.

Therefore EFI-EST provides the ability to generate SSNs using only the sequences of the domains specified by the input Pfam or InterPro family (*vide infra*) for the BLAST pairwise comparison/edge calculation. The following paragraphs describe this process.

Each of the fourteen protein databases integrated in the InterPro database define domain boundaries for the members of their families; when specified, EFI-EST trims full-length input sequences to the domain sequences specified by the input family. *SSNs generated using domains must be interpreted with caution*. First, domains can be split by insertions, including additional domains; EFI-EST recognizes each partial domain as a separate domain, uses each in the BLAST, and includes each as a separate node in the SSN. The haloalkanoate dehalogenase (HAD) superfamily is an example of this complication[7, 8]. Second, InterPro families/domains often are defined by more than one protein database; EFI-EST recognizes the domain boundaries defined by each database and generates a node for each database, usually with different domain boundaries, for the same sequence. *When an SSN is generated using domains, one input family should be used, either one Pfam family or one InterPro family/domain defined by one database.*

[The InterProScan (https://www.ebi.ac.uk/interpro/search/sequence-search) result for a sequence provides 1) a summary of the InterPro domains and families that are identified using the hidden Markov models (HMMs) for the fourteen protein family databases that are integrated in the InterPro database and 2) the domain boundaries that are identified by each database that contributes to an InterPro family/domain.]

When domains are used, the "**Dataset Analysis**" tab on the "**Dataset Completed**" page (Section II.E.2 for EFI-EST) provides the "**Sequence as a Function of Domain Length Histogram (First Step for Alignment Score Threshold Selection)**" for the trimmed domains for the full set of UniProt IDs in the input dataset, irrespective of whether the input dataset was generated with UniProt IDs, UniProt90 cluster IDs, or UniProt50 cluster IDs. This histogram should be used to evaluate the minimum length of the domain for selecting the minimum alignment score threshold used to generate the initial SSN; in analogy to the sequences in datasets for full-length ("untrimmed") proteins, the trimmed domain dataset likely will contain fragments that can be removed with the "**Sequence Length Restrictions Options**" on the "**SSN Finalization**" tab on "**Dataset Completed**" page.

When domains are used, the "**Dataset Analysis**" tab (Section II.E.2 for EFI-EST) also provides 1) the "**Sequence as a Function of Full-Length Histogram (UniProt IDs)**" for all full-length sequences in the input dataset (UniProt IDs), irrespective of whether the input dataset used UniProt IDs, UniProt90 cluster IDs, or UniProt50 cluster IDs, 2) the "**Sequence as a Function of Domain Length Histogram (UniRef Cluster IDs)**" for the trimmed domains for the UniRef cluster IDs in the input dataset if UniRef90 or UniRef50 clusters were used (the sequences of these domains are used by BLAST for the edge calculations), and 3) the "**Sequence as a Function of Full-Length Histogram (UniRef Cluster IDs)**" for the full-length distribution of the UniRef cluster IDs in the input dataset if UniRef90 or UniRef50 clusters were used. This complete set of full-length and domain length histograms provides a comprehensive description of the full-length and domain sequences in the input dataset.

When domains are used, irrespective of whether UniProt IDs, UniRef90 cluster IDs, or UniRef50 cluster IDs are used to generate the SSN, the "**Sequence as a Function of Domain Length Histogram (First Step for Alignment Score Threshold Selection)**" should be used for determining the alignment score threshold for generating the initial SSN. If UniProt IDs are used to generate the SSN, the domain sequences in "**Sequence as a Function of Domain Length Histogram (First Step for Alignment Score Threshold Selection)"** are used by BLAST for the edge calculation. If UniRef90 or UniRef50 cluster IDs are used to generate the SSN, the domain sequences in "**Sequence as a Function of Domain Length Histogram (UniRef Cluster IDs)**" are used by BLAST for the edge calculation.

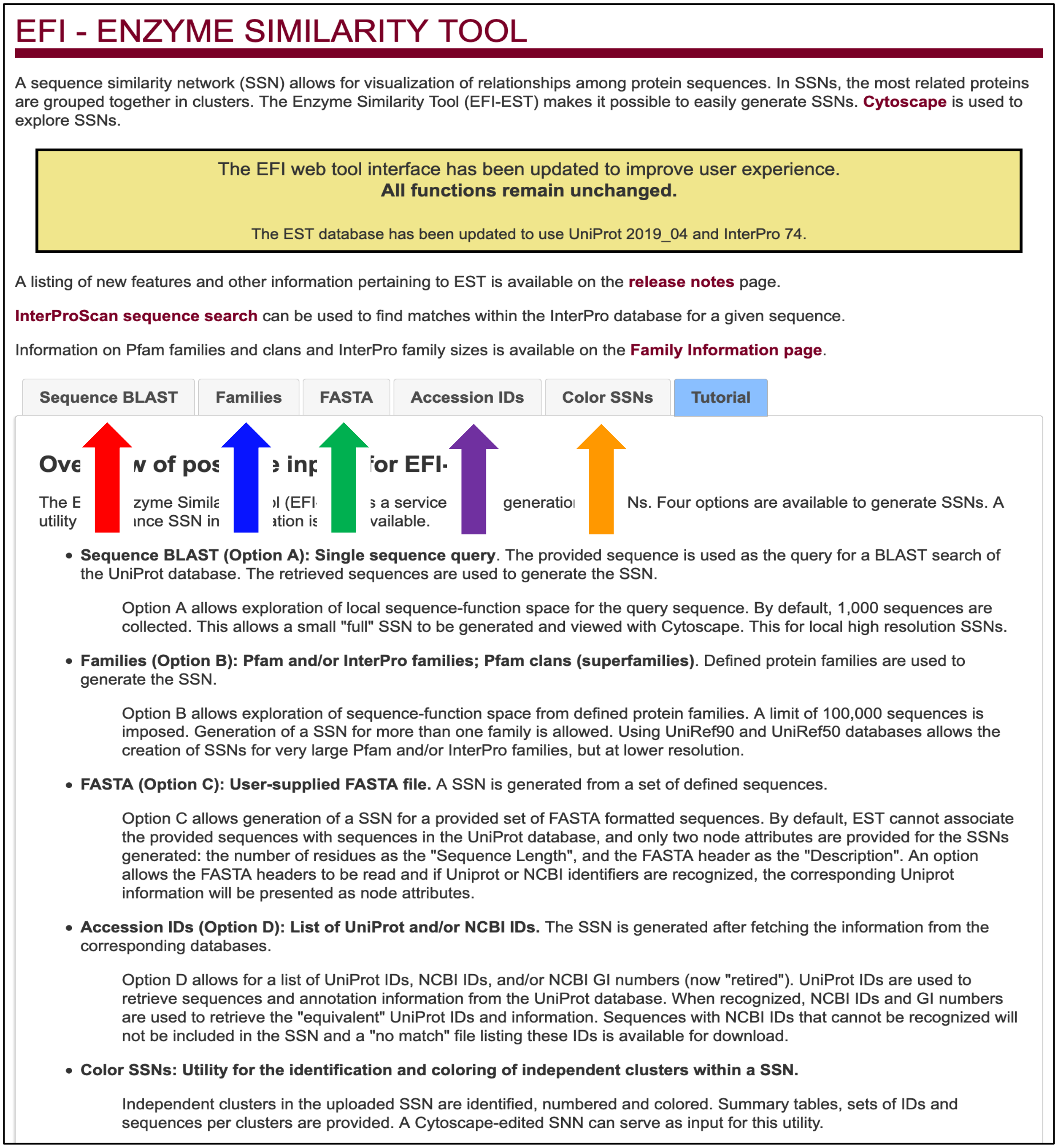
G. Computer Cluster

The web tools are hosted on a cluster purchased by the Enzyme Function Initiative (EFI; NIH U54GM093342) and maintained by the Computer Network Resource Group (CNRG) at the Institute for Genomic Biology (IGB) at the University of Illinois, Urbana-Champaign. The cluster includes 1) one webserver with 24 cores and 384 GB RAM; 2) 27 compute nodes, 24 with 24 cores and 384 GB RAM and three with 24 cores and 192 GB RAM; 3) one large memory node with 48 cores and 3TB RAM; 4) 164TB of dedicated storage; and 5) one high performance MySQL server for database storage. By default, the BLAST jobs for EFI-EST are assigned to 48 cores, the CGFP marker identification jobs are assigned to one compute node with 24 cores, and the CGFP metagenome abundance jobs are assigned to 24 cores. BLAST-reduce jobs (to aggregate and minimize the large dataset returned by BLAST) for EFI-EST and the generation of the GNNs by EFI-GNT (collection of genome neighborhood data) are assigned to the large memory node. *Access is provided without charge*.

II. EFI-EST (https://efi.igb.illinois.edu/efi-est/)

Sequence and Attribute Database for Generating SSNs. Sequences and bioinformatic information (for SSN node attributes; from the UniProtKB database) and ID mapping files (from the InterPro database; including an equivalence table for NCBI and UniProt IDs) are downloaded with each release of the InterPro database (every other release of the UniProt database; approximately every 8 weeks) to construct the local databases used by the tools.

Home Page. The EFI-EST home page provides access to four options for generating SSNs (Options A, B, C, and D; accessed by tabs; red, blue, green, and magenta arrows, respectively) that differ by how the input dataset is provided



EFI - ENZYME SIMILARITY TOOL

A sequence similarity network (SSN) allows for visualization of relationships among protein sequences. In SSNs, the most related proteins are grouped together in clusters. The Enzyme Similarity Tool (EFI-EST) makes it possible to easily generate SSNs. **Cytoscape** is used to explore SSNs.

The EFI web tool interface has been updated to improve user experience.
All functions remain unchanged.

The EST database has been updated to use UniProt 2019_04 and InterPro 74.

A listing of new features and other information pertaining to EST is available on the **release notes** page.

InterProScan sequence search can be used to find matches within the InterPro database for a given sequence.

Information on Pfam families and clans and InterPro family sizes is available on the **Family Information page**.

Ove w of pos e inp for EFI-

The E zyme Simila l (EFI s a service generatio Ns. Four options are available to generate SSNs. A utility ance SSN in ation is vailable.

- **Sequence BLAST (Option A): Single sequence query**. The provided sequence is used as the query for a BLAST search of the UniProt database. The retrieved sequences are used to generate the SSN.

  Option A allows exploration of local sequence-function space for the query sequence. By default, 1,000 sequences are collected. This allows a small "full" SSN to be generated and viewed with Cytoscape. This for local high resolution SSNs.

- **Families (Option B): Pfam and/or InterPro families; Pfam clans (superfamilies)**. Defined protein families are used to generate the SSN.

  Option B allows exploration of sequence-function space from defined protein families. A limit of 100,000 sequences is imposed. Generation of a SSN for more than one family is allowed. Using UniRef90 and UniRef50 databases allows the creation of SSNs for very large Pfam and/or InterPro families, but at lower resolution.

- **FASTA (Option C): User-supplied FASTA file.** A SSN is generated from a set of defined sequences.

  Option C allows generation of a SSN for a provided set of FASTA formatted sequences. By default, EST cannot associate the provided sequences with sequences in the UniProt database, and only two node attributes are provided for the SSNs generated: the number of residues as the "Sequence Length", and the FASTA header as the "Description". An option allows the FASTA headers to be read and if Uniprot or NCBI identifiers are recognized, the corresponding Uniprot information will be presented as node attributes.

- **Accession IDs (Option D): List of UniProt and/or NCBI IDs.** The SSN is generated after fetching the information from the corresponding databases.

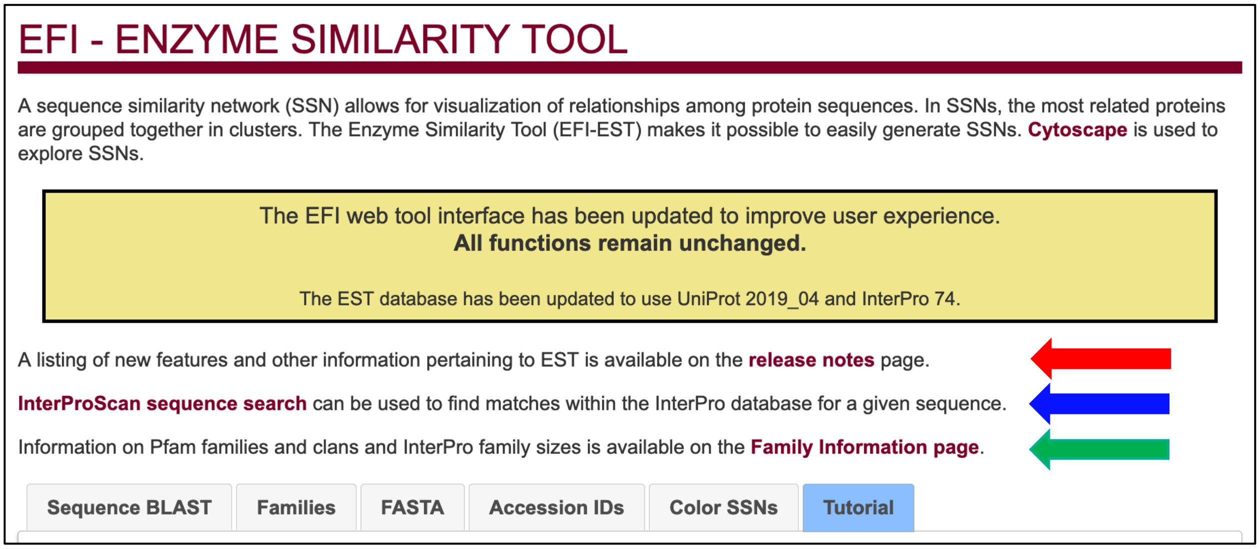
  Option D allows for a list of UniProt IDs, NCBI IDs, and/or NCBI GI numbers (now "retired"). UniProt IDs are used to retrieve sequences and annotation information from the UniProt database. When recognized, NCBI IDs and GI numbers are used to retrieve the "equivalent" UniProt IDs and information. Sequences with NCBI IDs that cannot be recognized will not be included in the SSN and a "no match" file listing these IDs is available for download.

- **Color SSNs: Utility for the identification and coloring of independent clusters within a SSN.**

  Independent clusters in the uploaded SSN are identified, numbered and colored. Summary tables, sets of IDs and sequences per clusters are provided. A Cytoscape-edited SNN can serve as input for this utility.

The Home Page also provides tab for a Color SSNs utility (orange arrow) for assigning unique colors to each cluster in an input SSN as well as unique numbers to each cluster and singleton in the SSN. Some options for changing input parameters are the shared among the input Options but complete descriptions are provided for each Option.

The top of the home page has links to 1) "**release notes**" that provide descriptions of the enhancements/updates made to EFI-EST (red arrow), 2) "**InterProScan sequence**" search to identify the InterPro and Pfam protein family databases that contain an input sequence (blue arrow), and 3) "**Family Information page**" (green arrow) that provides information about the sizes of each of the InterPro and Pfam families as well as links to the InterPro/Pfam family page.


EFI - ENZYME SIMILARITY TOOL

A sequence similarity network (SSN) allows for visualization of relationships among protein sequences. In SSNs, the most related proteins are grouped together in clusters. The Enzyme Similarity Tool (EFI-EST) makes it possible to easily generate SSNs. **Cytoscape** is used to explore SSNs.

The EFI web tool interface has been updated to improve user experience.
All functions remain unchanged.

The EST database has been updated to use UniProt 2019_04 and InterPro 74.

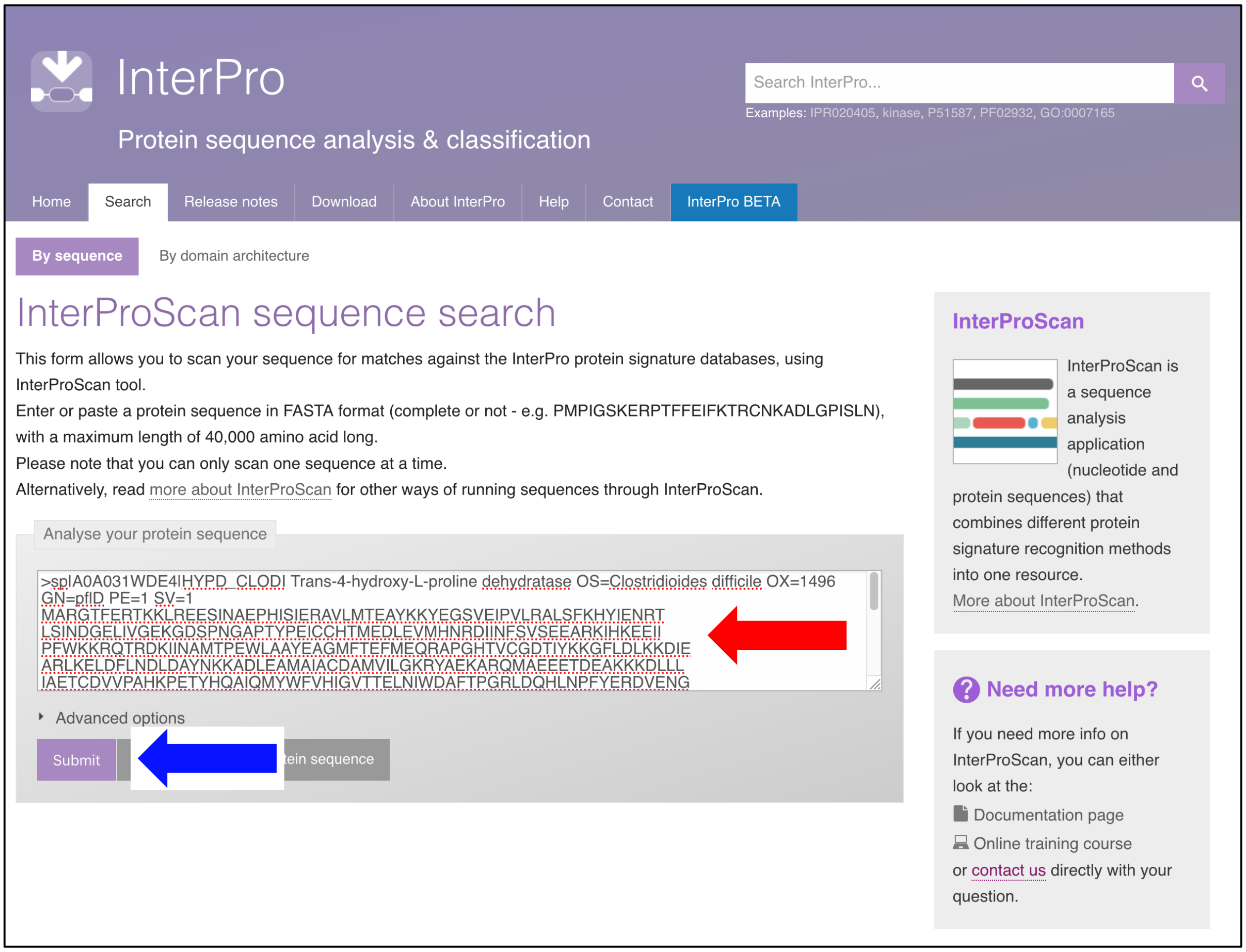
A listing of new features and other information pertaining to EST is available on the **release notes** page.

InterProScan sequence search can be used to find matches within the InterPro database for a given sequence.

Information on Pfam families and clans and InterPro family sizes is available on the **Family Information page**.

Sequence BLAST (Options A), **FASTA** (Option C), and **Accession IDs** (Option D) allow the user-defined input sequences to be placed in the context of an InterPro or Pfam family in the generated SSN; **Families** (Option B) allows a user to generate an SSN for specified InterPro and/or Pfam families. The "**InterProScan sequence**" link to the InterProScan page (https://www.ebi.ac.uk/interpro/search/sequence-search) provided by EMBL/EBI allows the user to quickly identify the InterPro (and Pfam) families that include the input sequence; the "**Family Information page**" link to the EFI-EST Family Information page (https://efi.igb.illinois.edu/efi-est/family_list.php) provides summaries of the sizes of all of the InterPro and Pfam families (UniProt IDs, UniRef90 cluster IDs, and UniRef50 cluster IDs) that will be useful in the selection of input parameters for generating SSNs (*vide infra*).

InterProScan. The InterProScan page provides a box for entering a sequence (with or without FASTA header; red arrow). The sequence is submitted by clicking on the "Submit" button (blue arrow).



InterProScan sequence search

This form allows you to scan your sequence for matches against the InterPro protein signature databases, using InterProScan tool.

Enter or paste a protein sequence in FASTA format (complete or not - e.g. PMPIGSKERPTFFEIFKTRCNKADLGPISLN), with a maximum length of 40,000 amino acid long.

Please note that you can only scan one sequence at a time.

Alternatively, read more about InterProScan for other ways of running sequences through InterProScan.

Analyse your protein sequence

```
>spIA0A031WDE4IHYPD_CLODI Trans-4-hydroxy-L-proline dehydratase OS=Clostridioides difficile OX=1496
GN=pflD PE=1 SV=1
MARGTFERTKKLREESINAEPHISIERAVLMTEAYKKYEGSVEIPVLRALSFKHYIENRT
LSINDGELIVGEKGDSPNGAPTYPEICCHTMEDLEVMHNRDIINFSVSEEARKIHKEEII
PFWKKRQTRDKIINAMTPEWLAAYEAGMFTEFMEQRAPGHTVCGDTIYKKGFLDLKKDIE
ARLKELDFLNDLDAYNKKADLEAMAIACDAMVILGKRYAEKARQMAEEETDEAKKKDLLL
IAETCDVVPAHKPETYHQAIQMYWFVHIGVTTELNIWDAFTPGRLDQHLNPFYERDVENG
```

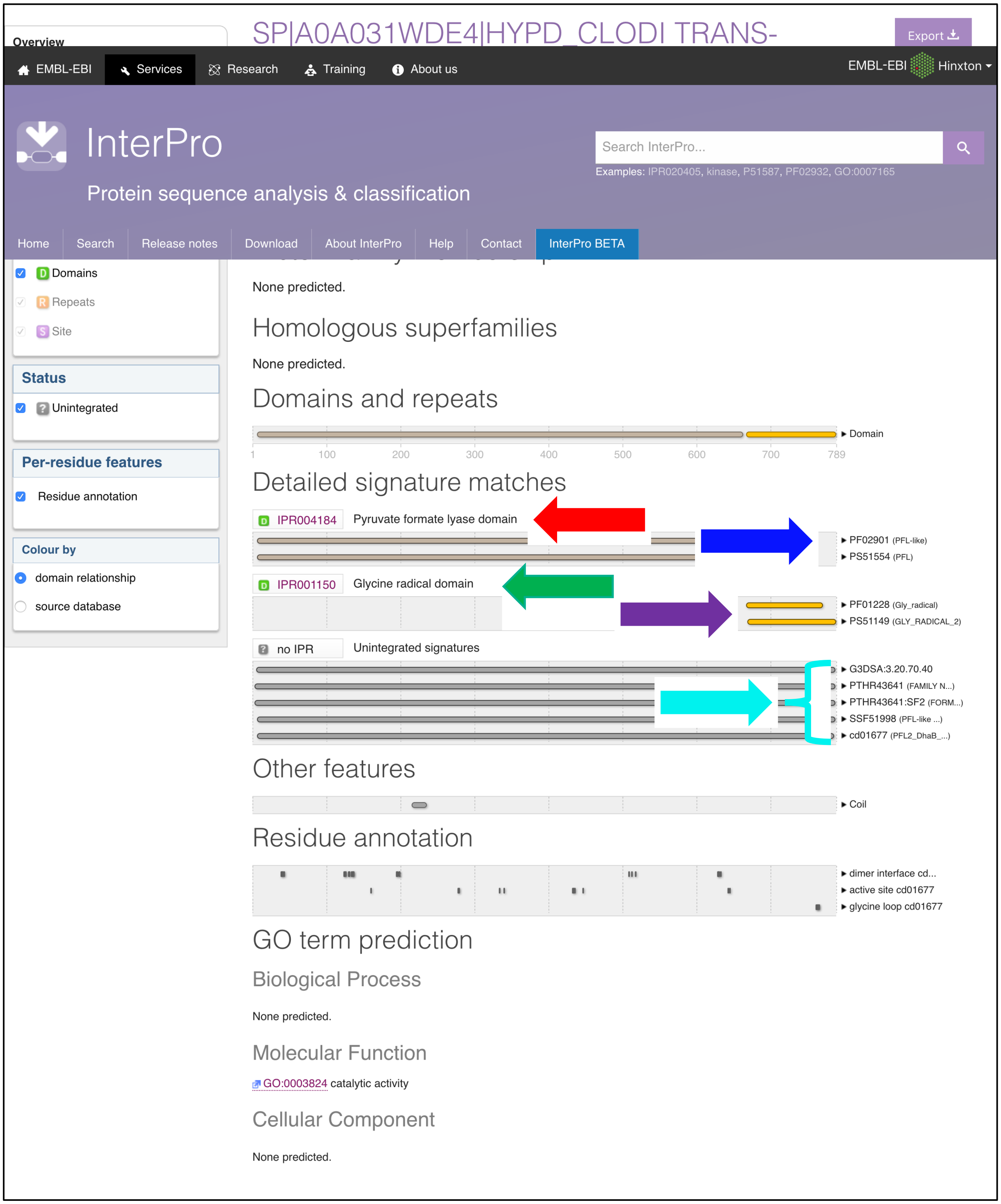
InterProScan

InterProScan is a sequence analysis application (nucleotide and protein sequences) that combines different protein signature recognition methods into one resource.
More about InterProScan.

Need more help?

If you need more info on InterProScan, you can either look at the:
Documentation page
Online training course
or contact us directly with your question.

The results are returned after a short wait.



SP|A0A031WDE4|HYPD_CLODI TRANS-

☑ D Domains

☐ R Repeats

☐ S Site

Status

☑ ? Unintegrated

Per-residue features

☑ Residue annotation

Colour by

● domain relationship

○ source database

None predicted.

Homologous superfamilies

None predicted.

Domains and repeats

▸ Domain

Detailed signature matches

D IPR004184 Pyruvate formate lyase domain

▸ PF02901 (PFL-like)
▸ PS51554 (PFL)

D IPR001150 Glycine radical domain

▸ PF01228 (Gly_radical)
▸ PS51149 (GLY_RADICAL_2)

? no IPR Unintegrated signatures

▸ G3DSA:3.20.70.40
▸ PTHR43641 (FAMILY N...)
▸ PTHR43641:SF2 (FORM...)
▸ SSF51998 (PFL-like ...)
▸ cd01677 (PFL2_DhaB_...)

Other features

▸ Coil

Residue annotation

▸ dimer interface cd...
▸ active site cd01677
▸ glycine loop cd01677

GO term prediction

Biological Process

None predicted.

Molecular Function

GO:0003824 catalytic activity

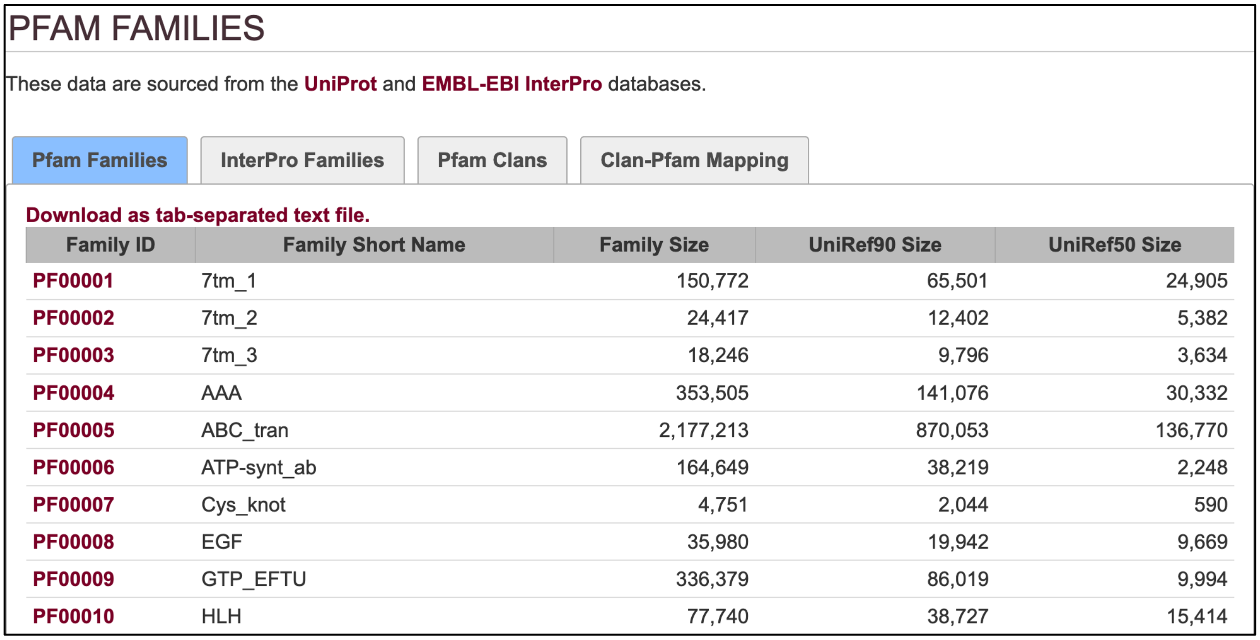
Cellular Component

None predicted.

In this example, the input sequence, UniProt ID A0A031WDE4, is a member of the Pyruvate formate lyase InterPro domain (IPR004184; red arrow); this domain is determined (blue arrow) by Pfam family PF02901 (PFL-like) and, also, PROSITE family PS51554 (PFL2). IPR004184 is also referred to as the “glycyl radical enzyme (GRE) superfamily”; it is used as the example in the article text.

InterProScan also identified the Glycine radical InterPro domain (IPR001150; green arrow) that is determined (magenta arrow) by Pfam family PF01228 (Gly_radical) and PROSITE family PS51149 (GLY_RADICAL_2). In addition, InterProScan identified families defined by several other its component databases (cyan arrow/bracket), although these were not incorporated into an InterPro family or domain.

Family Information Page. This page provides tables of the UniProt IDs, UniRef90 cluster IDs, and UniRef50 cluster IDs in each Pfam family, InterPro family, and Pfam clan.



PFAM FAMILIES

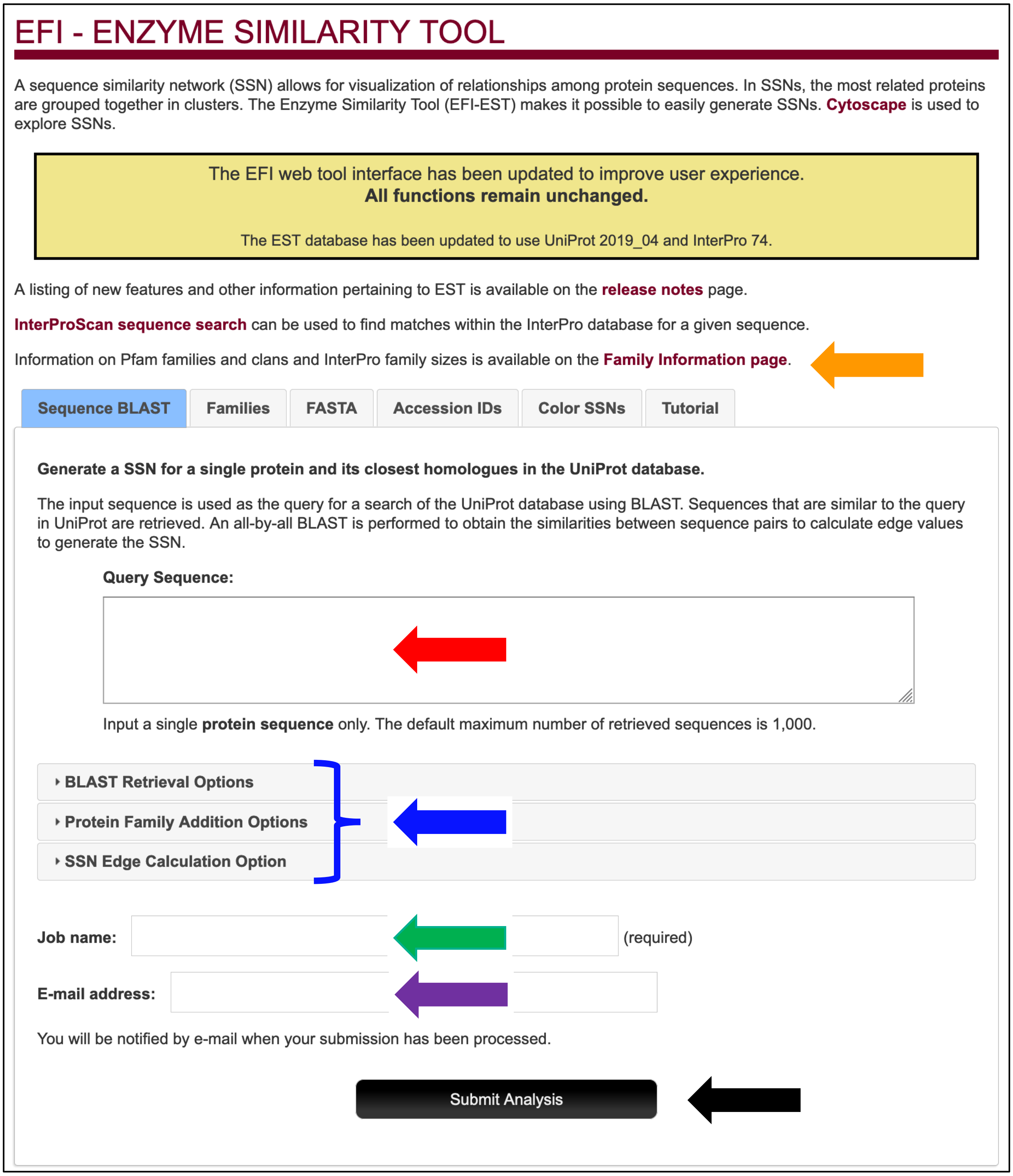
These data are sourced from the **UniProt** and **EMBL-EBI InterPro** databases.

Download as tab-separated text file.

Family ID	Family Short Name	Family Size	UniRef90 Size	UniRef50 Size
PF00001	7tm_1	150,772	65,501	24,905
PF00002	7tm_2	24,417	12,402	5,382
PF00003	7tm_3	18,246	9,796	3,634
PF00004	AAA	353,505	141,076	30,332
PF00005	ABC_tran	2,177,213	870,053	136,770
PF00006	ATP-synt_ab	164,649	38,219	2,248
PF00007	Cys_knot	4,751	2,044	590
PF00008	EGF	35,980	19,942	9,669
PF00009	GTP_EFTU	336,379	86,019	9,994
PF00010	HLH	77,740	38,727	15,414

The tables are accessed with the “**Pfam Families**”, “**InterPro Families**”, and “**Pfam Clans**” tabs. In these tables, the “**Family ID**” is a link to the Pfam or InterPro database page describing the family, domain, or clan. The fourth table accessed with the “**Clan-Pfam Mapping**” tab details the Pfam clan-family mapping (Pfam clans are groups of homologous Pfam families; membership in Pfam families is not mutually exclusive so the same accession ID can be a member of multiple Pfam families). Each of these tables can be downloaded as a tab-delimited text file.

A. EFI-EST Home Page, “Sequence BLAST” tab: Single sequence query (Option A)



EFI - ENZYME SIMILARITY TOOL

A sequence similarity network (SSN) allows for visualization of relationships among protein sequences. In SSNs, the most related proteins are grouped together in clusters. The Enzyme Similarity Tool (EFI-EST) makes it possible to easily generate SSNs. **Cytoscape** is used to explore SSNs.

The EFI web tool interface has been updated to improve user experience.
All functions remain unchanged.

The EST database has been updated to use UniProt 2019_04 and InterPro 74.

A listing of new features and other information pertaining to EST is available on the **release notes** page.

InterProScan sequence search can be used to find matches within the InterPro database for a given sequence.

Information on Pfam families and clans and InterPro family sizes is available on the **Family Information page**.

Generate a SSN for a single protein and its closest homologues in the UniProt database.

The input sequence is used as the query for a search of the UniProt database using BLAST. Sequences that are similar to the query in UniProt are retrieved. 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.

Job name: (required)

E-mail address:

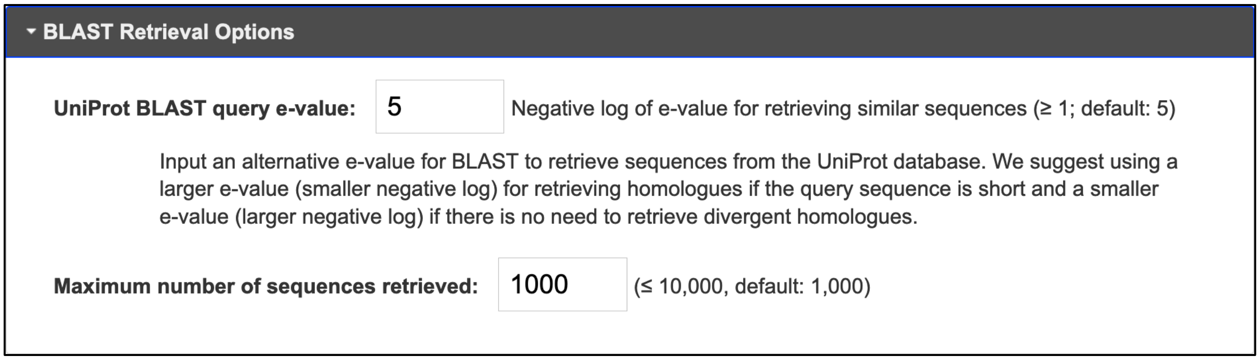
You will be notified by e-mail when your submission has been processed.

Query Entry. A user-provided sequence (with or without a FASTA header) for a BLAST search of the UniProt database is entered in the "**Query Sequence**" box (red arrow); the sequences retrieved by the query are used to generate the SSN. By default, 1,000 homologues are collected because this number should allow a full SSN (all accession IDs) to be opened and viewed with Cytoscape on most computers; a larger number of sequences (≤10,000) can be specified in the "**BLAST Retrieval Options**".

In the SSN, the "name" and "shared name" node attributes for the query are "zINPUTSEQ". The "Sequence Source" node attribute identifies the origin of the sequence: "INPUT" identifies the query; "BLASTHIT" identifies the homologues identified by BLAST. For the input sequence, the "Description" node attribute is "Input Sequence"; for the homologues collected by BLAST, the full set of node attributes is provided.

The default parameters can be changed using the options in the accordion windows below the box for the input sequence (blue arrow and bracket): "**BLAST Retrieval Options**", "**Protein Family Addition Options**" and "**SSN Edge Calculation Options**".

BLAST Retrieval Options. The default e-value (negative base-10 log) used to collect homologues is 5; this can be changed (≥0, integer) by entering an integer in the "**UniProt BLAST query E-value**" box. If the search sequence is short, a value <5 is recommended. Also, the "**Maximum number of sequences retrieved**" can be increased (default 1,000; ≤10,000).



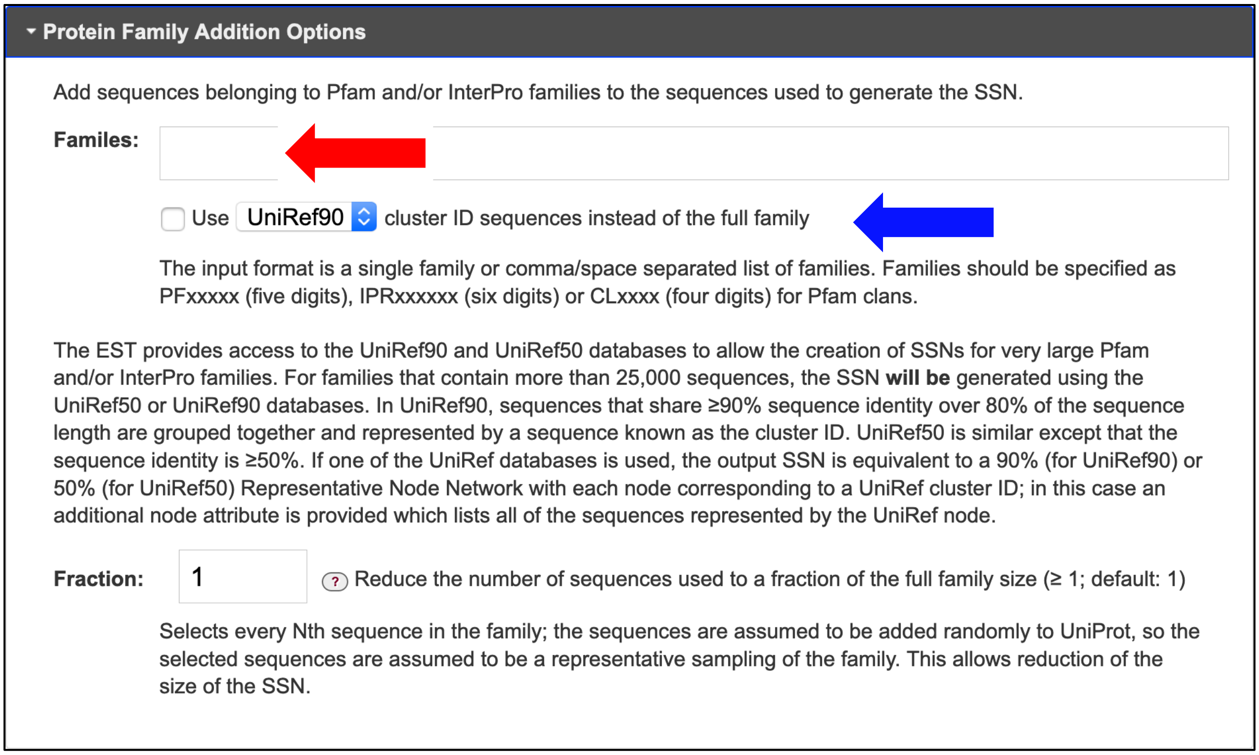
▾ BLAST Retrieval Options

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

Input an alternative e-value for BLAST to retrieve sequences from the UniProt database. We suggest using a larger e-value (smaller negative log) for retrieving homologues if the query sequence is short and a smaller e-value (larger negative log) if there is no need to retrieve divergent homologues.

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

Protein Family Addition Options. The sequences identified by BLAST can be added to one or more Pfam and/or InterPro family(ies) and/or Pfam clans so that the BLAST hits can be placed in the context of an entire protein family. The family/class ID is entered in the "**Families**" box (red arrow): Pfam family, **PFnnnnn** (five numbers); InterPro family, **IPRnnnnnn** (six numbers); Pfam clan, **CLnnnn** (four numbers).



Add sequences belonging to Pfam and/or InterPro families to the sequences used to generate the SSN.

Familes:

□ Use UniRef90 cluster ID sequences instead of the full family

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.

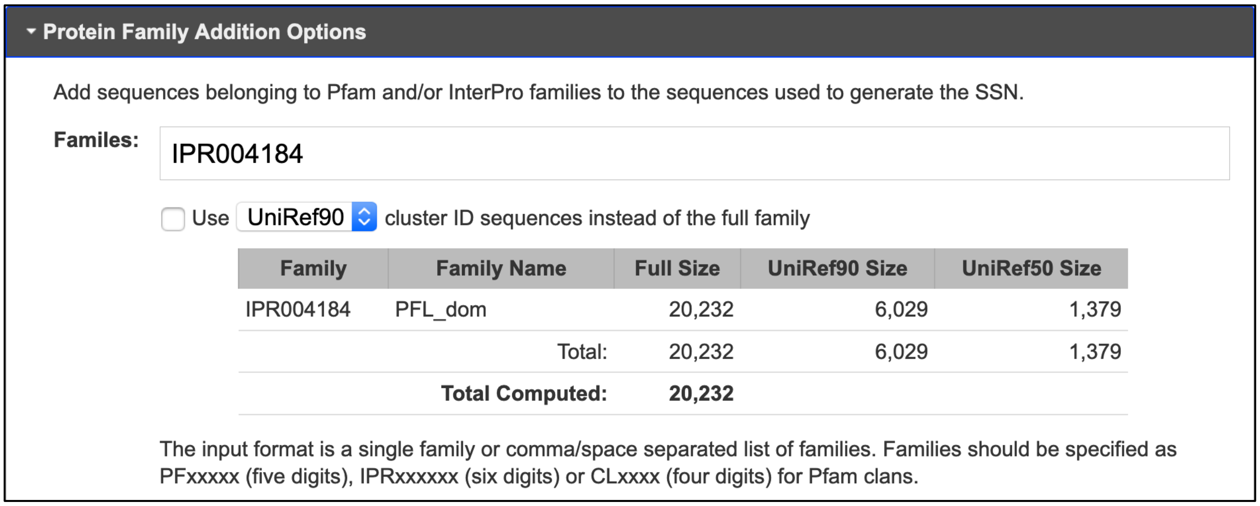
The EST provides access to the UniRef90 and UniRef50 databases to allow the creation of SSNs for very large Pfam and/or InterPro families. For families that contain more than 25,000 sequences, the SSN **will be** generated using the UniRef50 or UniRef90 databases. In UniRef90, sequences that share ≥90% sequence identity over 80% of the sequence length are grouped together and represented by a sequence known as the cluster ID. UniRef50 is similar except that the sequence identity is ≥50%. If one of the UniRef databases is used, the output SSN is equivalent to a 90% (for UniRef90) or 50% (for UniRef50) Representative Node Network with each node corresponding to a UniRef cluster ID; in this case an additional node attribute is provided which lists all of the sequences represented by the UniRef node.

Fraction: 1 ? Reduce the number of sequences used to a fraction of the full family size (≥ 1; default: 1)

Selects every Nth sequence in the family; the sequences are assumed to be added randomly to UniProt, so the selected sequences are assumed to be a representative sampling of the family. This allows reduction of the size of the SSN.

Depending on the number of sequences in the family (*vide infra* for restrictions on the total number of sequences in the input dataset), UniProt IDs, UniRef90 cluster IDs, or UniRef50 cluster IDs can be selected using the check box and UniRef90/UniRef50 menu (blue arrow in the "**Protein Family Additions Option**" accordion).

For example, addition of a UniProt family (IPR004184, the glycyl radical enzyme superfamily example used in the main text):



▾ Protein Family Addition Options

Add sequences belonging to Pfam and/or InterPro families to the sequences used to generate the SSN.

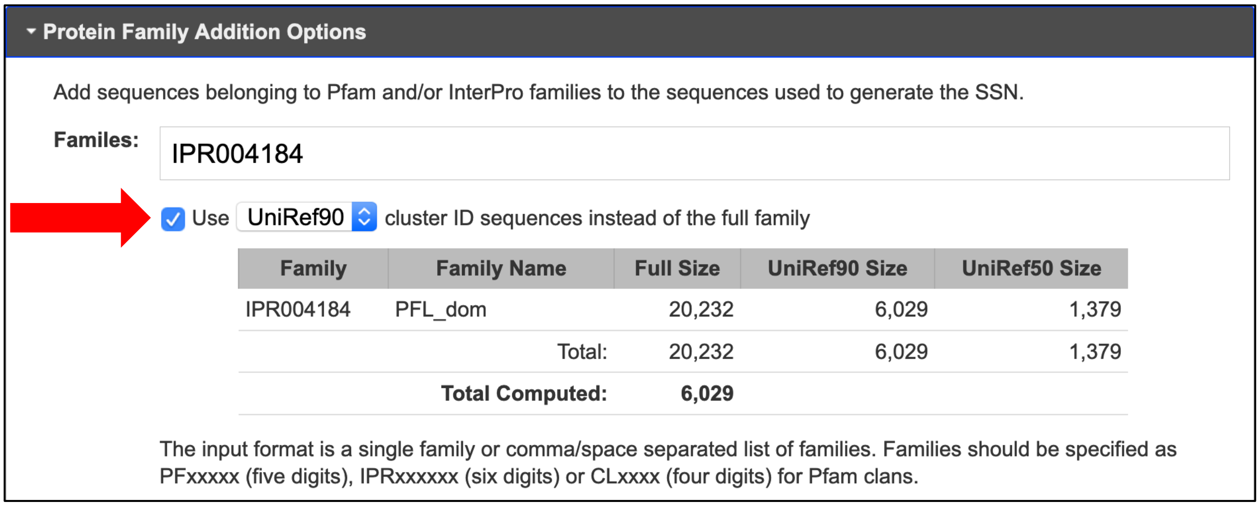
Familes: IPR004184

□ Use UniRef90 cluster ID sequences instead of the full family

Family	Family Name	Full Size	UniRef90 Size	UniRef50 Size
IPR004184	PFL_dom	20,232	6,029	1,379
	Total:	20,232	6,029	1,379
	Total Computed:	**20,232**		

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.

If either UniRef90 or UniRef50 cluster IDs are selected/imposed (red arrow), the user must confirm the selection before the job is submitted (blue arrow).



Protein Family Addition Options

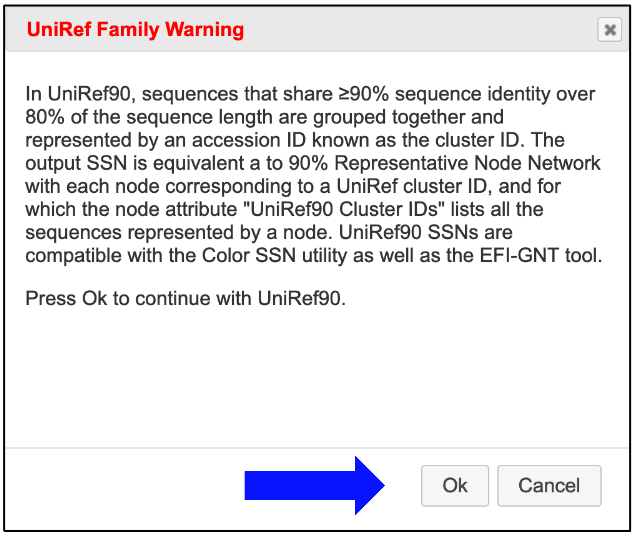
Add sequences belonging to Pfam and/or InterPro families to the sequences used to generate the SSN.

Familes: IPR004184

☑ Use UniRef90 cluster ID sequences instead of the full family

Family	Family Name	Full Size	UniRef90 Size	UniRef50 Size
IPR004184	PFL_dom	20,232	6,029	1,379
	Total:	20,232	6,029	1,379
	Total Computed:	**6,029**		

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.

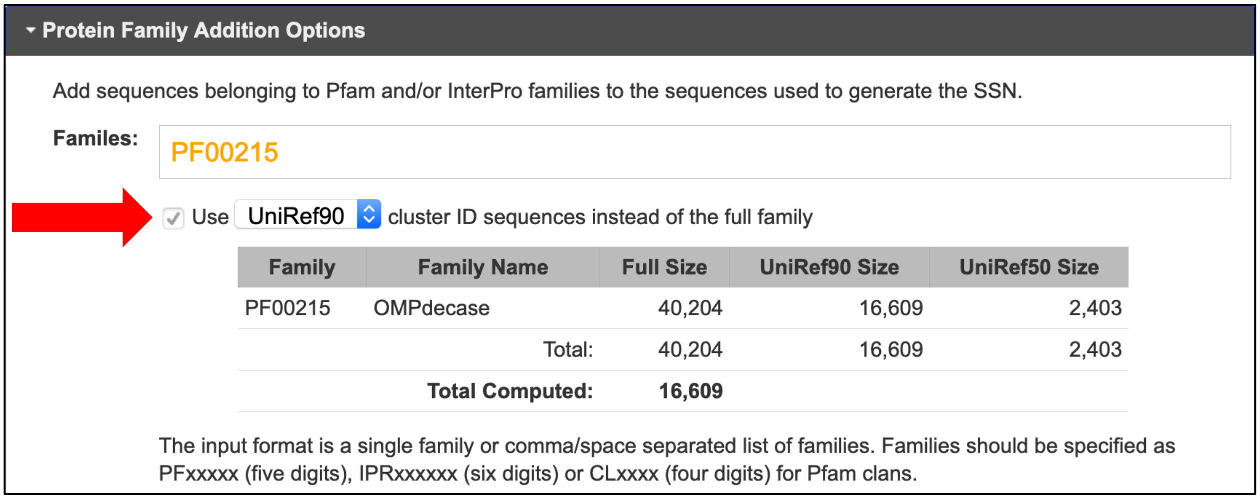


UniRef Family Warning

In UniRef90, sequences that share ≥90% sequence identity over 80% of the sequence length are grouped together and represented by an accession ID known as the cluster ID. The output SSN is equivalent a to 90% Representative Node Network with each node corresponding to a UniRef cluster ID, and for which the node attribute "UniRef90 Cluster IDs" lists all the sequences represented by a node. UniRef90 SSNs are compatible with the Color SSN utility as well as the EFI-GNT tool.

Press Ok to continue with UniRef90.

For a family with >25,000 sequences, UniRef90 cluster IDs are required and will be imposed if the number of UniRef90 clusters is ≤100,000. The user must confirm the selection before the job is submitted (blue arrow).



▾ Protein Family Addition Options

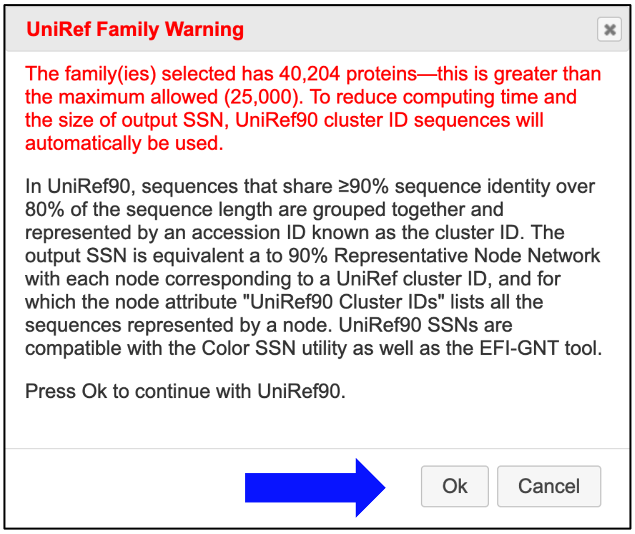
Add sequences belonging to Pfam and/or InterPro families to the sequences used to generate the SSN.

Familes: PF00215

☑ Use UniRef90 cluster ID sequences instead of the full family

Family	Family Name	Full Size	UniRef90 Size	UniRef50 Size
PF00215	OMPdecase	40,204	16,609	2,403
	Total:	40,204	16,609	2,403
	Total Computed:	**16,609**		

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.



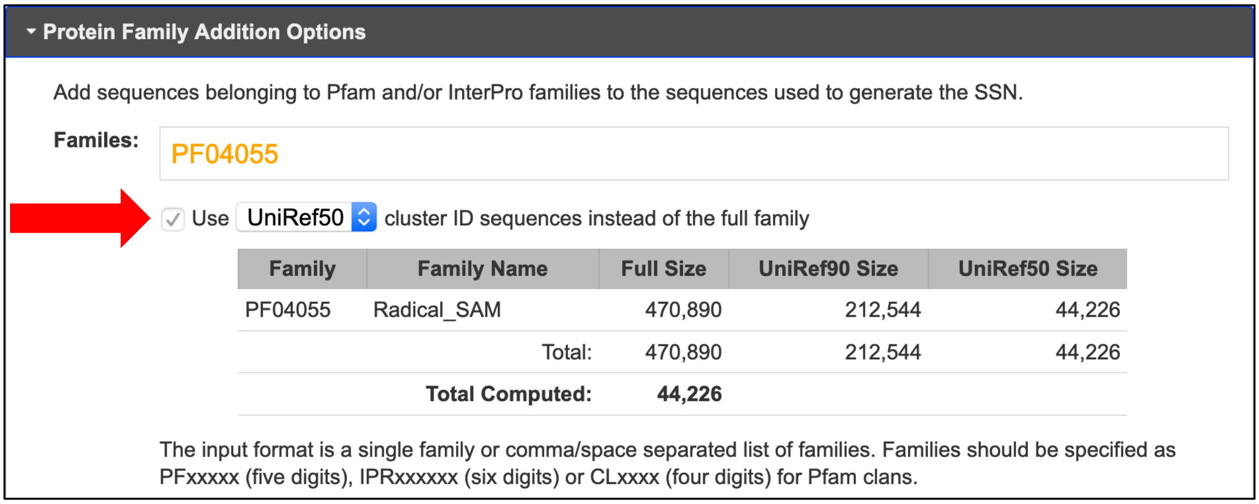
UniRef Family Warning

The family(ies) selected has 40,204 proteins—this is greater than the maximum allowed (25,000). To reduce computing time and the size of output SSN, UniRef90 cluster ID sequences will automatically be used.

In UniRef90, sequences that share ≥90% sequence identity over 80% of the sequence length are grouped together and represented by an accession ID known as the cluster ID. The output SSN is equivalent a to 90% Representative Node Network with each node corresponding to a UniRef cluster ID, and for which the node attribute "UniRef90 Cluster IDs" lists all the sequences represented by a node. UniRef90 SSNs are compatible with the Color SSN utility as well as the EFI-GNT tool.

Press Ok to continue with UniRef90.

For families with >100,000 UniRef90 cluster IDs, UniRef50 cluster IDs are required and will be imposed if the number of UniRef50 clusters is ≤100,000. The user must confirm the selection before the job is submitted (blue arrow).



Protein Family Addition Options

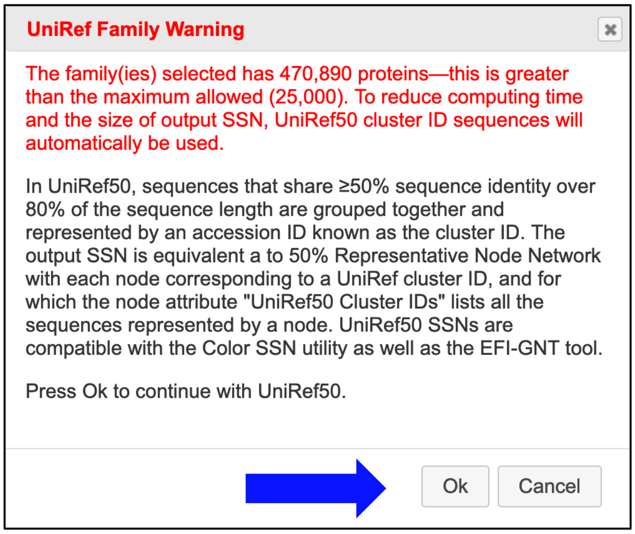
Add sequences belonging to Pfam and/or InterPro families to the sequences used to generate the SSN.

Familes: PF04055

☑ Use UniRef50 cluster ID sequences instead of the full family

Family	Family Name	Full Size	UniRef90 Size	UniRef50 Size
PF04055	Radical_SAM	470,890	212,544	44,226
	Total:	470,890	212,544	44,226
	Total Computed:	**44,226**		

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.



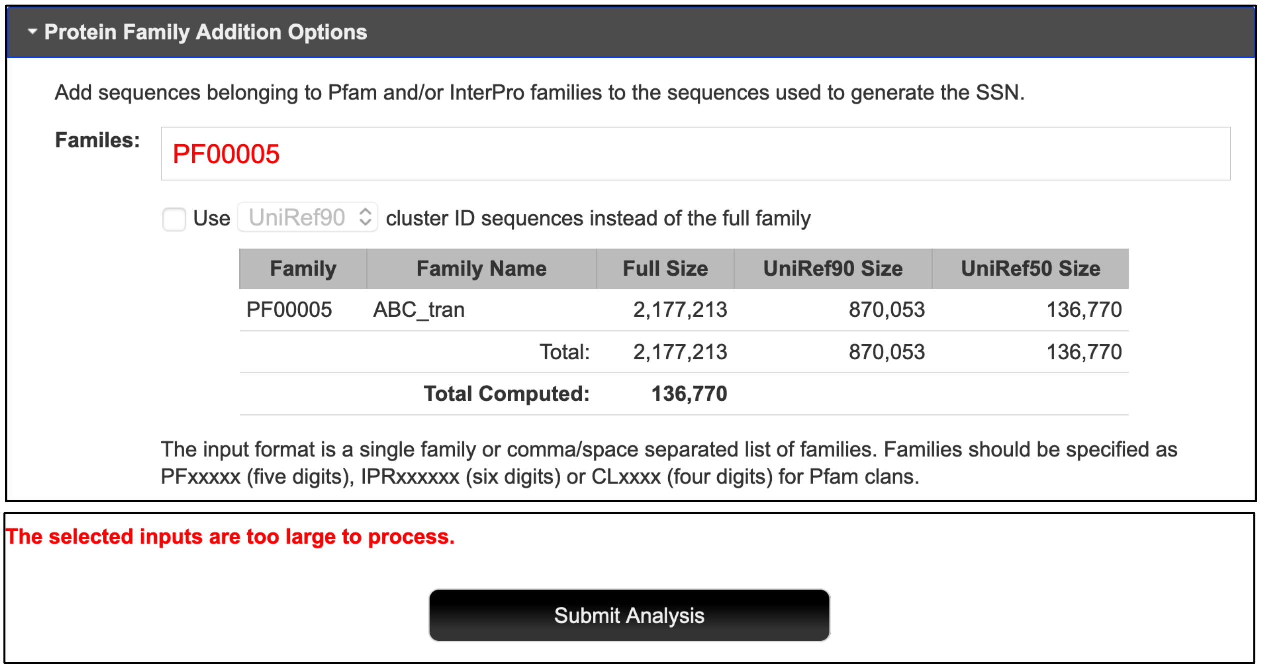
UniRef Family Warning

The family(ies) selected has 470,890 proteins—this is greater than the maximum allowed (25,000). To reduce computing time and the size of output SSN, UniRef50 cluster ID sequences will automatically be used.

In UniRef50, sequences that share ≥50% sequence identity over 80% of the sequence length are grouped together and represented by an accession ID known as the cluster ID. The output SSN is equivalent a to 50% Representative Node Network with each node corresponding to a UniRef cluster ID, and for which the node attribute "UniRef50 Cluster IDs" lists all the sequences represented by a node. UniRef50 SSNs are compatible with the Color SSN utility as well as the EFI-GNT tool.

Press Ok to continue with UniRef50.

A maximum of 100,000 UniRef50 clusters is imposed—the job cannot be submitted when the “Submit Analysis” button is clicked.



▾ Protein Family Addition Options

Add sequences belonging to Pfam and/or InterPro families to the sequences used to generate the SSN.

Familes: PF00005

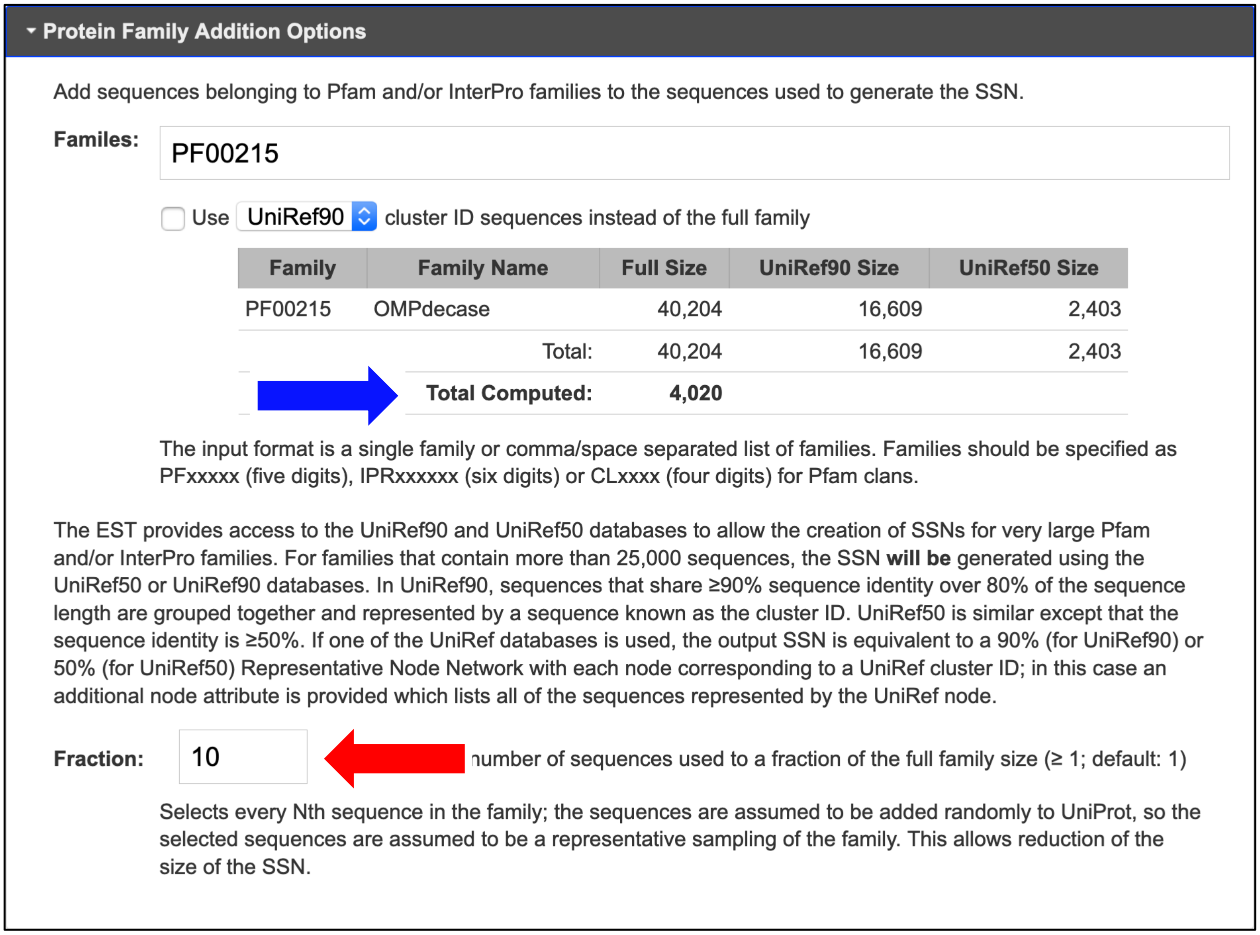
□ Use UniRef90 cluster ID sequences instead of the full family

Family	Family Name	Full Size	UniRef90 Size	UniRef50 Size
PF00005	ABC_tran	2,177,213	870,053	136,770
	Total:	2,177,213	870,053	136,770
	Total Computed:	**136,770**		

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.

The selected inputs are too large to process.

The user also can specify that a representative fraction of the UniProt IDs in the Pfam and/or InterPro family(ies) be added to the BLAST hits: an integer (N) is entered in the "**Fraction**" box (red arrow), with EFI-EST selecting every Nth sequence in the family; the sequences are assumed to be added randomly to UniProt, so the selected sequences are assumed to be a representative sampling of the family (blue arrow). Using a family fraction allows the BLAST hits to be placed in the context of a large family. When using the Fraction option, SwissProt-curated proteins are always included. *The fraction option and UniRef90 or UniRef50 cluster IDs cannot be used together*.



Protein Family Addition Options

Add sequences belonging to Pfam and/or InterPro families to the sequences used to generate the SSN.

Familes: PF00215

□ Use UniRef90 cluster ID sequences instead of the full family

Family	Family Name	Full Size	UniRef90 Size	UniRef50 Size
PF00215	OMPdecase	40,204	16,609	2,403
	Total:	40,204	16,609	2,403
	Total Computed:	**4,020**		

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.

The EST provides access to the UniRef90 and UniRef50 databases to allow the creation of SSNs for very large Pfam and/or InterPro families. For families that contain more than 25,000 sequences, the SSN **will be** generated using the UniRef50 or UniRef90 databases. In UniRef90, sequences that share ≥90% sequence identity over 80% of the sequence length are grouped together and represented by a sequence known as the cluster ID. UniRef50 is similar except that the sequence identity is ≥50%. If one of the UniRef databases is used, the output SSN is equivalent to a 90% (for UniRef90) or 50% (for UniRef50) Representative Node Network with each node corresponding to a UniRef cluster ID; in this case an additional node attribute is provided which lists all of the sequences represented by the UniRef node.

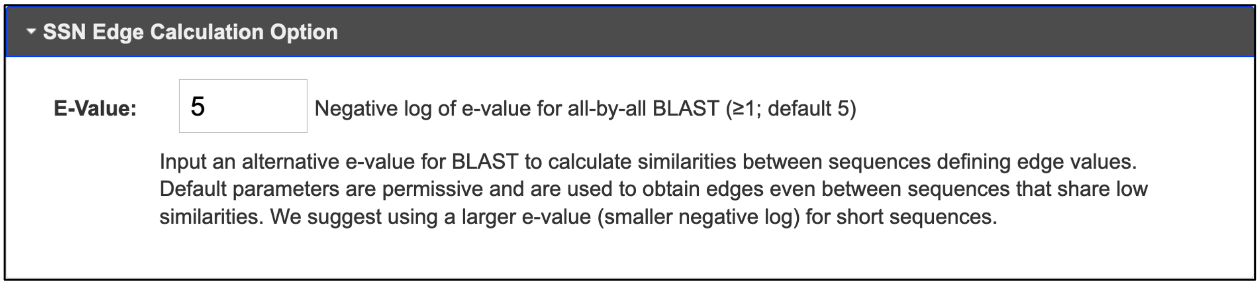
Fraction: 10 ıumber of sequences used to a fraction of the full family size (≥ 1; default: 1)

Selects every Nth sequence in the family; the sequences are assumed to be added randomly to UniProt, so the selected sequences are assumed to be a representative sampling of the family. This allows reduction of the size of the SSN.

In the SSN, the “Sequence Source” node attribute “FAMILY” identifies the sequences in added Pfam and/or InterPro family(ies). If a family is added using UniProt IDs and the UniProt IDs for a “BLASTHIT” sequence and a “FAMILY” sequence are identical, the full SSN will contain a merged node with “FAMILY+BLASTHIT” as the “Sequence Source”; the “User IDs in Cluster” node attribute also will identify the “BLASTHIT” ID. For a rep node SSN, the “Sequence Source” node attribute for the metanode will be “FAMILY+BLASTHIT” if it contains one or more merged “FAMILY+BLASTHIT” nodes; the “User IDs in Cluster” node attribute will be a list of “BLASTHIT” IDs in the metanode. (In the rep node SSNs, the “name”/”shared name” of the metanode is the ID of the longest sequence in the metanode.)

If a family is added using UniRef clusters and the accession ID of the BLASTHIT is the same as either the UniRef cluster ID or a UniProt ID contained in the UniRef cluster, the full SSN will contain a merged node with “FAMILY+BLASTHIT” identifying the “Sequence Source”; the “User IDs in Cluster” node attribute will be a list of BLASTHIT IDs in the node. For a rep node SSN, the “Sequence Source” node attribute for the metanode will be “FAMILY+BLASTHIT” if it contains one or more merged “FAMILY+BLASTHIT” nodes; the “User IDs in Cluster” node attribute will be a list of “BLASTHIT” IDs in the metanode. (In the rep node SSNs, the “name”/”shared name” of the metanode is the ID of the longest cluster ID in the metanode.)

SSN Edge Calculation Option. The user can select an alternative e-value (negative base-10 log) for calculating the edges by entering an integer in the "**E-Value**" box (default is 5). If the sequences in the dataset are short, a value <5 should be entered. A larger value might be selected for calculating the edges if the alignment score to generate the SSN is known—there is no need to calculate an alignment score smaller than the alignment score used to generate the SSN (the negative base-10 log of an e-value and the alignment score are similar in magnitude).



E-Value: 5 Negative log of e-value for all-by-all BLAST (≥1; default 5)

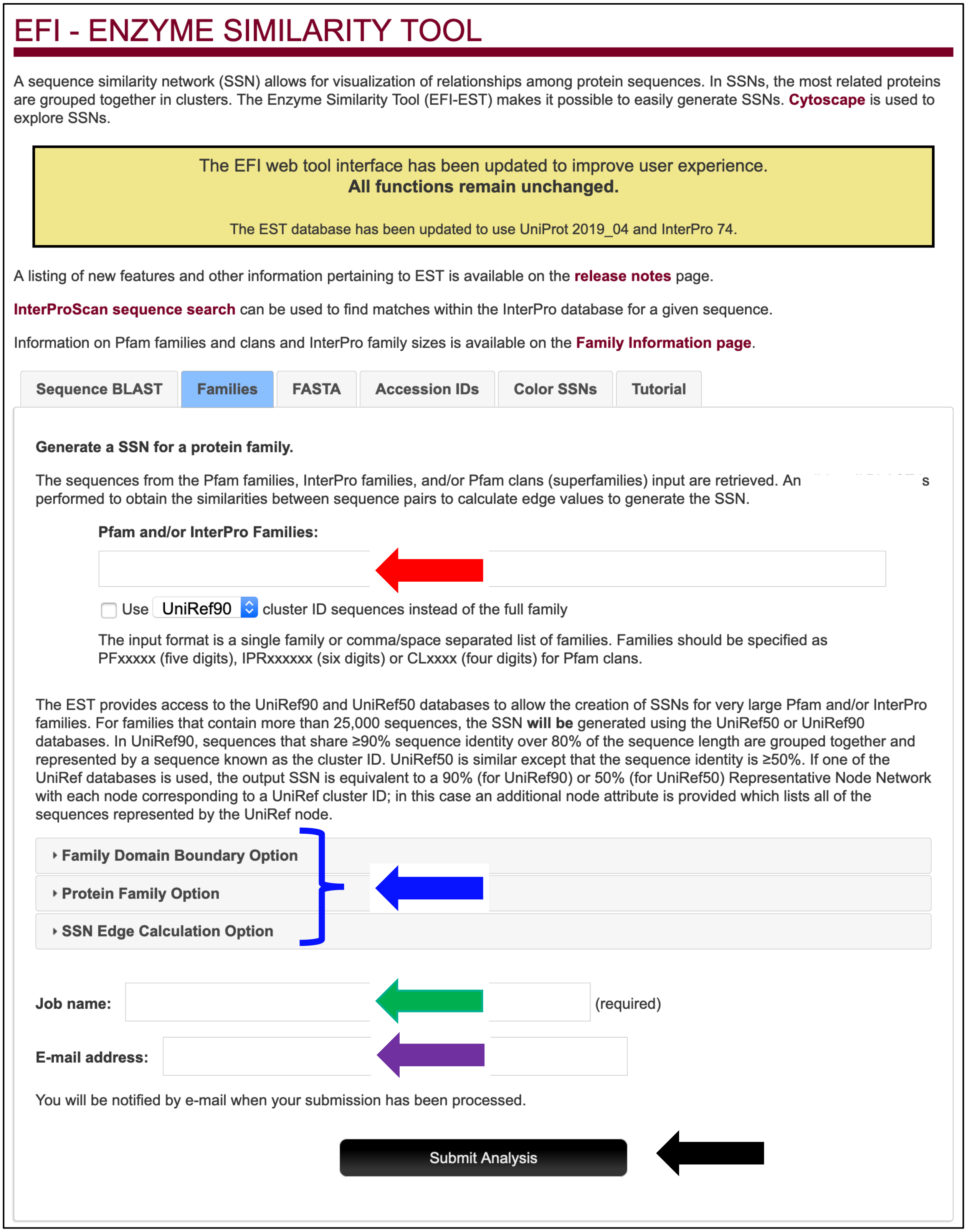
Input an alternative e-value for BLAST to calculate similarities between sequences defining edge values. Default parameters are permissive and are used to obtain edges even between sequences that share low similarities. We suggest using a larger e-value (smaller negative log) for short sequences.

A "**Job name**" (green arrow) is required that will be used on the "**Previous Jobs**" panel on the EFI-EST home page and the "**Job History**" page. On the "**Previous Jobs**" panel and the "**Job History**" page, these jobs are designated "**Sequence BLAST**".

An "**E-mail address**" (magenta arrow) is required. The job is initiated with the "**Submit Analysis**" button (black arrow).

The progress of the job is available on the "**Previous Jobs**" panel: "**PENDING**" when the job is received, "**RUNNING**" while the job is running, and a link to the results page when the job is finished. Also, an e-mail is sent to the user when the job is started; a second e-mail is sent when the job is finished that includes a link to the results page.

B. EFI-EST Home Page, “Families” tab: Pfam family(ies), InterPro family(ies), and/or Pfam clan(s) (Option B)



EFI - ENZYME SIMILARITY TOOL

A sequence similarity network (SSN) allows for visualization of relationships among protein sequences. In SSNs, the most related proteins are grouped together in clusters. The Enzyme Similarity Tool (EFI-EST) makes it possible to easily generate SSNs. **Cytoscape** is used to explore SSNs.

The EFI web tool interface has been updated to improve user experience.
All functions remain unchanged.

The EST database has been updated to use UniProt 2019_04 and InterPro 74.

A listing of new features and other information pertaining to EST is available on the **release notes** page.

InterProScan sequence search can be used to find matches within the InterPro database for a given sequence.

Information on Pfam families and clans and InterPro family sizes is available on the **Family Information page**.

Generate a SSN for a protein family.

The sequences from the Pfam families, InterPro families, and/or Pfam clans (superfamilies) input are retrieved. An [illegible] s performed to obtain the similarities between sequence pairs to calculate edge values to generate the SSN.

Pfam and/or InterPro Families:

□ Use UniRef90 cluster ID sequences instead of the full family

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.

The EST provides access to the UniRef90 and UniRef50 databases to allow the creation of SSNs for very large Pfam and/or InterPro families. For families that contain more than 25,000 sequences, the SSN **will be** generated using the UniRef50 or UniRef90 databases. In UniRef90, sequences that share ≥90% sequence identity over 80% of the sequence length are grouped together and represented by a sequence known as the cluster ID. UniRef50 is similar except that the sequence identity is ≥50%. If one of the UniRef databases is used, the output SSN is equivalent to a 90% (for UniRef90) or 50% (for UniRef50) Representative Node Network with each node corresponding to a UniRef cluster ID; in this case an additional node attribute is provided which lists all of the sequences represented by the UniRef node.

‣ Family Domain Boundary Option
‣ Protein Family Option
‣ SSN Edge Calculation Option

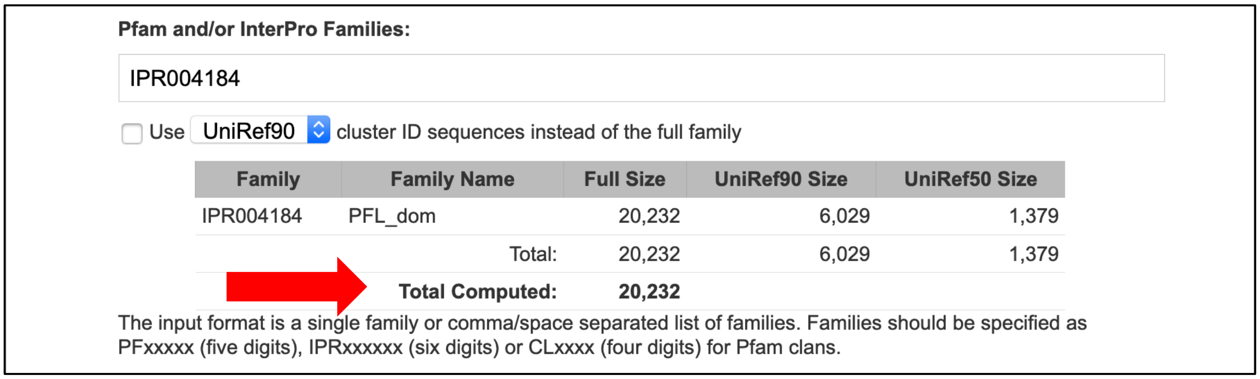
Job name: (required)

E-mail address:

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

Option B allows exploration of the sequence-function space in a protein family.

Family Entry. The SSN is generated using the sequences in one or more user-specified Pfam and/or InterPro family(ies) and/or Pfam clans; the family/class identifier is entered in the "**Pfam and/or InterPro Families**" box [Pfam family, **PFnnnnn** (five numbers); InterPro family, **IPRnnnnnn** (six numbers); Pfam clan, **CLnnnn** (four numbers)] (red arrow).



Pfam and/or InterPro Families:

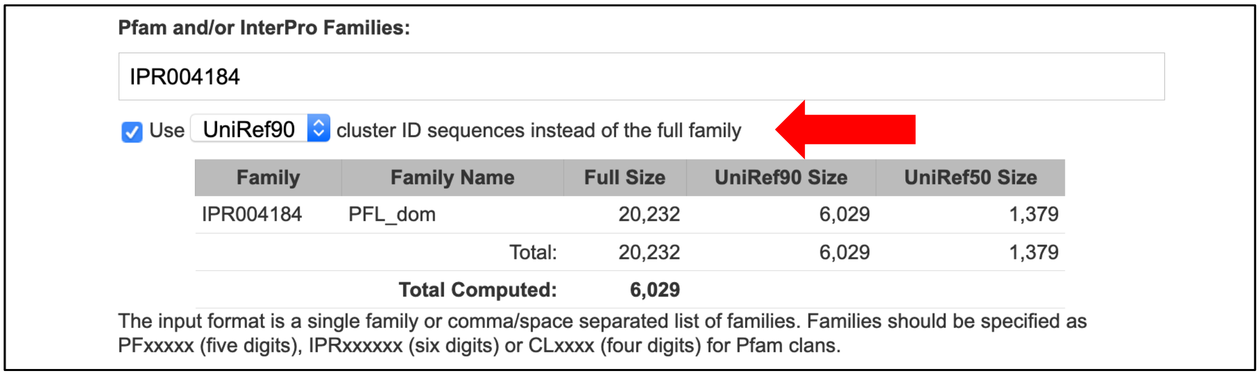
IPR004184

□ Use UniRef90 cluster ID sequences instead of the full family

Family	Family Name	Full Size	UniRef90 Size	UniRef50 Size
IPR004184	PFL_dom	20,232	6,029	1,379
	Total:	20,232	6,029	1,379
	Total Computed:	**20,232**		

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.

If either UniRef90 or UniRef50 cluster IDs are selected/imposed (red arrow), the user must confirm the selection before the job is submitted (blue arrow).



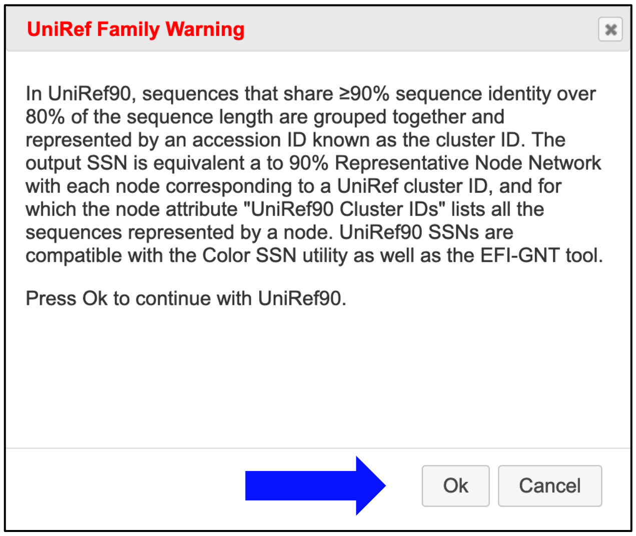
Pfam and/or InterPro Families:

IPR004184

☑ Use UniRef90 cluster ID sequences instead of the full family

Family	Family Name	Full Size	UniRef90 Size	UniRef50 Size
IPR004184	PFL_dom	20,232	6,029	1,379
	Total:	20,232	6,029	1,379
	Total Computed:	**6,029**		

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.

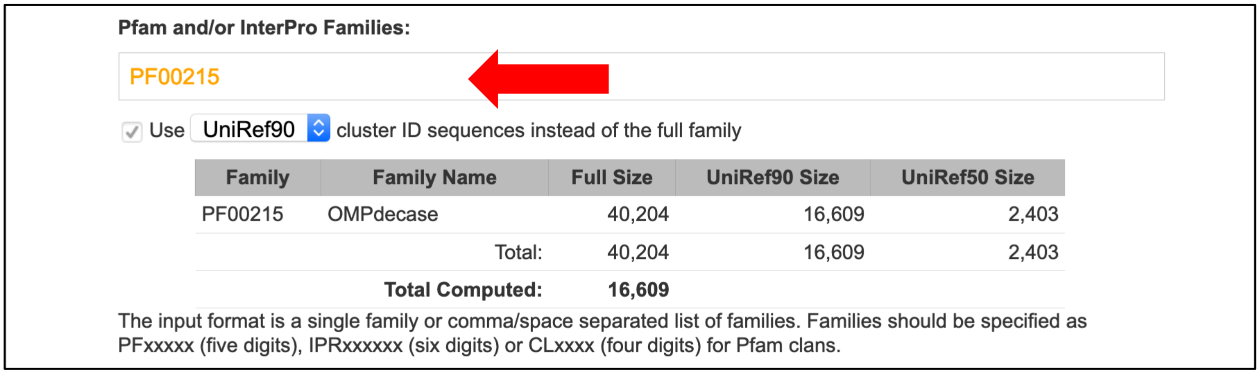


UniRef Family Warning

In UniRef90, sequences that share ≥90% sequence identity over 80% of the sequence length are grouped together and represented by an accession ID known as the cluster ID. The output SSN is equivalent a to 90% Representative Node Network with each node corresponding to a UniRef cluster ID, and for which the node attribute "UniRef90 Cluster IDs" lists all the sequences represented by a node. UniRef90 SSNs are compatible with the Color SSN utility as well as the EFI-GNT tool.

Press Ok to continue with UniRef90.

For a family with >25,000 sequences (red arrow), UniRef90 cluster IDs are required and will be imposed if the number of UniRef90 clusters is ≤100,000. The user must confirm the selection before the job is submitted (blue arrow).



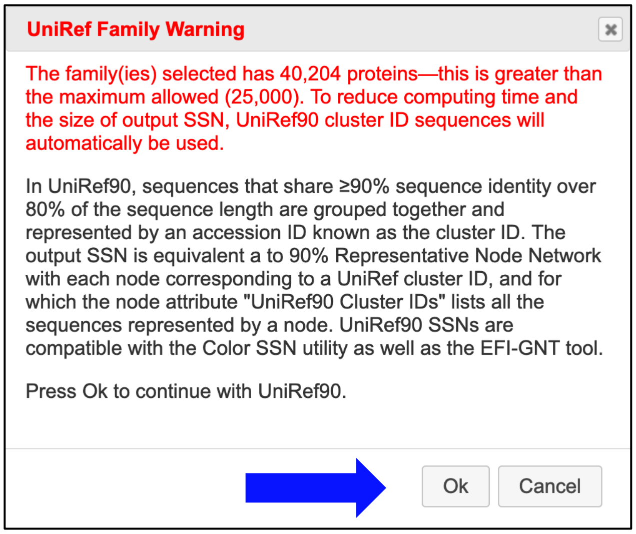
Pfam and/or InterPro Families:

PF00215

☑ Use UniRef90 cluster ID sequences instead of the full family

Family	Family Name	Full Size	UniRef90 Size	UniRef50 Size
PF00215	OMPdecase	40,204	16,609	2,403
	Total:	40,204	16,609	2,403
	Total Computed:	**16,609**		

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.



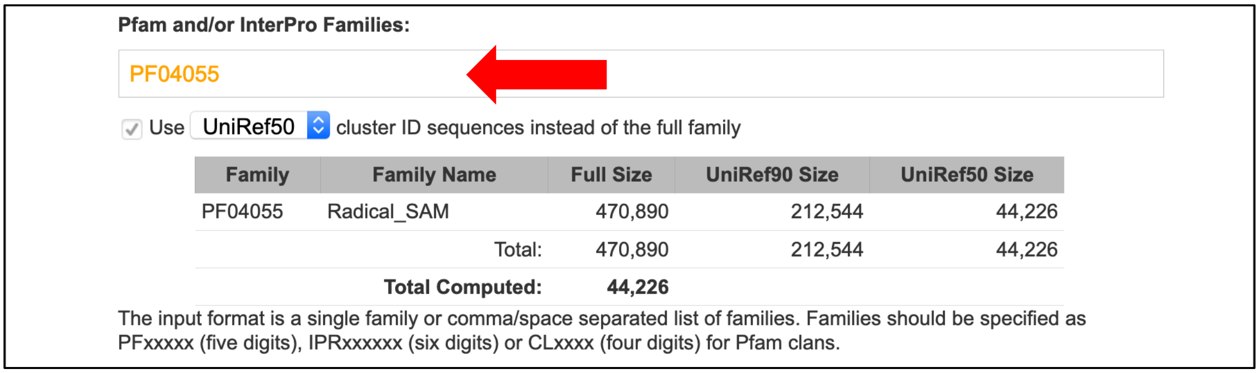
UniRef Family Warning

The family(ies) selected has 40,204 proteins—this is greater than the maximum allowed (25,000). To reduce computing time and the size of output SSN, UniRef90 cluster ID sequences will automatically be used.

In UniRef90, sequences that share ≥90% sequence identity over 80% of the sequence length are grouped together and represented by an accession ID known as the cluster ID. The output SSN is equivalent a to 90% Representative Node Network with each node corresponding to a UniRef cluster ID, and for which the node attribute "UniRef90 Cluster IDs" lists all the sequences represented by a node. UniRef90 SSNs are compatible with the Color SSN utility as well as the EFI-GNT tool.

Press Ok to continue with UniRef90.

For families with >100,000 UniRef90 cluster IDs (red arrow), UniRef50 cluster IDs are required and will be imposed if the number of UniRef50 clusters is ≤100,000. The user must confirm the selection before the job is submitted (blue arrow).



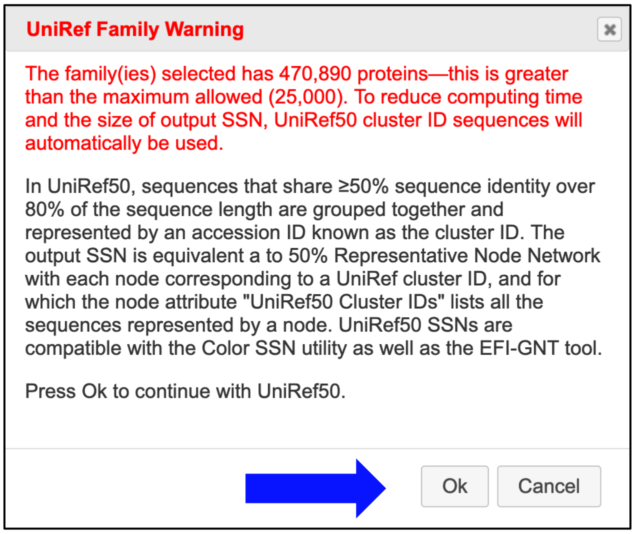
Pfam and/or InterPro Families:

PF04055

☑ Use UniRef50 cluster ID sequences instead of the full family

Family	Family Name	Full Size	UniRef90 Size	UniRef50 Size
PF04055	Radical_SAM	470,890	212,544	44,226
	Total:	470,890	212,544	44,226
	Total Computed:	**44,226**		

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.



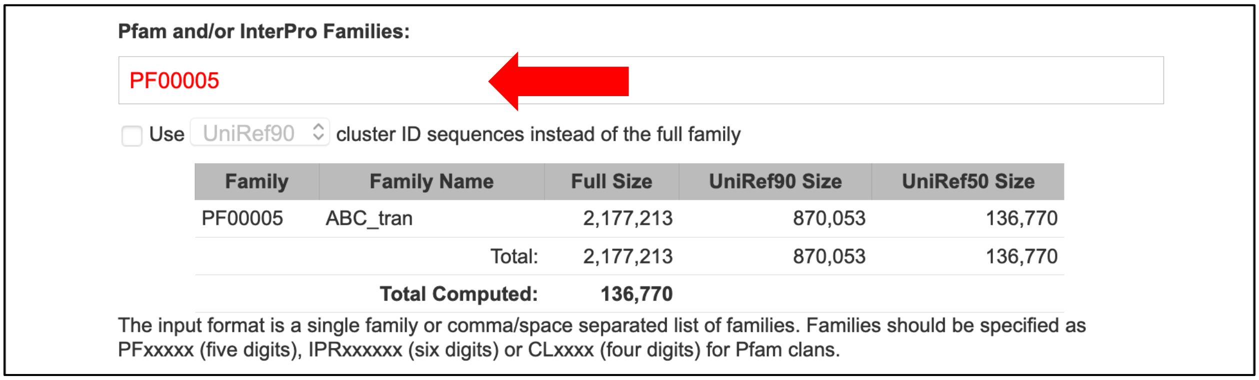
UniRef Family Warning

The family(ies) selected has 470,890 proteins—this is greater than the maximum allowed (25,000). To reduce computing time and the size of output SSN, UniRef50 cluster ID sequences will automatically be used.

In UniRef50, sequences that share ≥50% sequence identity over 80% of the sequence length are grouped together and represented by an accession ID known as the cluster ID. The output SSN is equivalent a to 50% Representative Node Network with each node corresponding to a UniRef cluster ID, and for which the node attribute "UniRef50 Cluster IDs" lists all the sequences represented by a node. UniRef50 SSNs are compatible with the Color SSN utility as well as the EFI-GNT tool.

Press Ok to continue with UniRef50.

A maximum of 100,000 UniRef50 clusters is imposed—the job cannot be submitted when the "Submit Analysis" button is clicked.



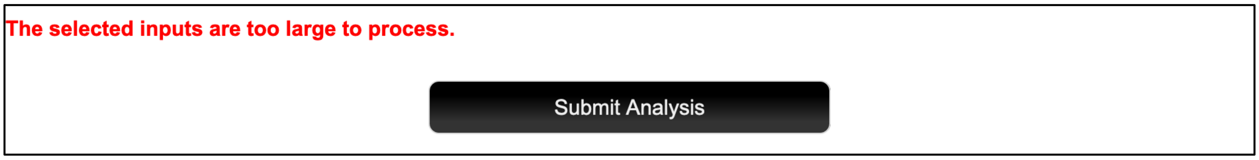
Pfam and/or InterPro Families:

PF00005

□ Use UniRef90 cluster ID sequences instead of the full family

Family	Family Name	Full Size	UniRef90 Size	UniRef50 Size
PF00005	ABC_tran	2,177,213	870,053	136,770
	Total:	2,177,213	870,053	136,770
	Total Computed:	**136,770**		

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.



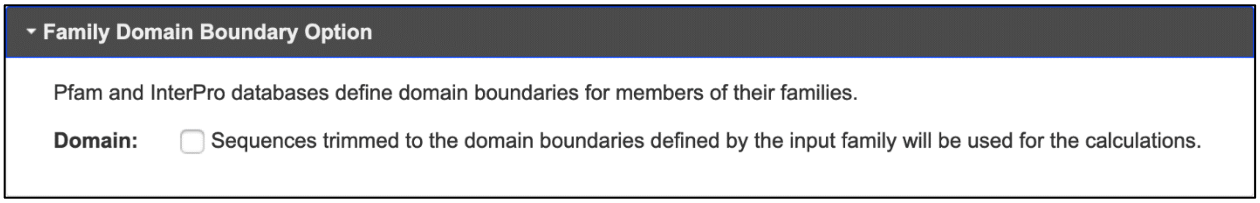
The selected inputs are too large to process.

In UniProt Release 2019_04, 1,666 of the 17,929 Pfam families include >25,000 UniProt IDs; 68 Pfam families include >100,000 UniRef90 clusters; and six Pfam families include >100,000 UniRef50 clusters [PF02518 (HATPase_c), PF00072 (Response_reg), PF00069 (Pkinase), PF00005 (ABC_Tran), PF00512 (HisKA), and PF07690 (MFS_1)]. Users interested in generating SSNs for these "extreme" families should contact the authors.

SSNs constructed using UniRef90 clusters are "high" resolution—the clusters are expected to be orthologues. SSNs constructed from UniRef50 clusters may be lower resolution resolution—the sequences may not be orthologues; however, for large protein families, these SSNs offer the advantage that the SSN files can be opened on laptop/desktop computers.

The default parameters can be changed using the options in the accordion windows below the box for the input sequence (blue arrow and bracket): "**Family Domain Boundaries Option**", "**Protein Family Option**", and "**SSN Edge Calculation Option**".

Family Domain Boundaries Option. By checking the box, Option B generates the SSN using sequences of the domains defined by the input family (for UniProt SSNs, all accession IDs; for UniRef SSNs, the cluster IDs).

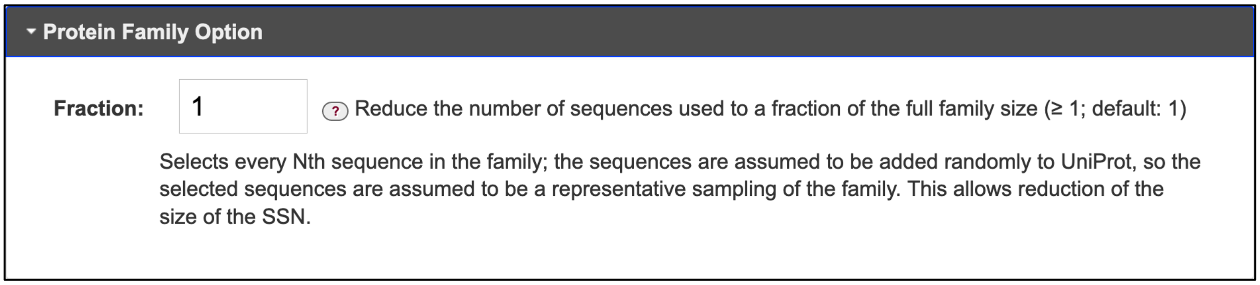


Pfam and InterPro databases define domain boundaries for members of their families.

Domain: □ Sequences trimmed to the domain boundaries defined by the input family will be used for the calculations.

An earlier section described both the advantages and disadvantages associated with generating SSNs from domains. Also, *when domains are used, the input should be a single Pfam family or an InterPro family/domain that is defined by one family database*. In the SSN, the "name" and "shared name" node attributes include the UniProt ID followed by colon-delimited residue numbers for N- and C-termini of the domain: UniProtID:N-terminus:C-terminus.

Protein Family Option. The user can specify that a representative fraction of the UniProt IDs in the user-specified Pfam and/or InterPro family(ies) be used to generate the SSN: an integer (N) is entered in the "**Fraction**" box, with EFI-EST selecting every Nth sequence in the family; the sequences are assumed to be added randomly to UniProt, so the selected sequences are assumed to be a representative sampling of the family. When using the Fraction option, SwissProt-curated proteins are always included. *The fraction option and UniRef90 or UniRef50 cluster IDs cannot be used together*.

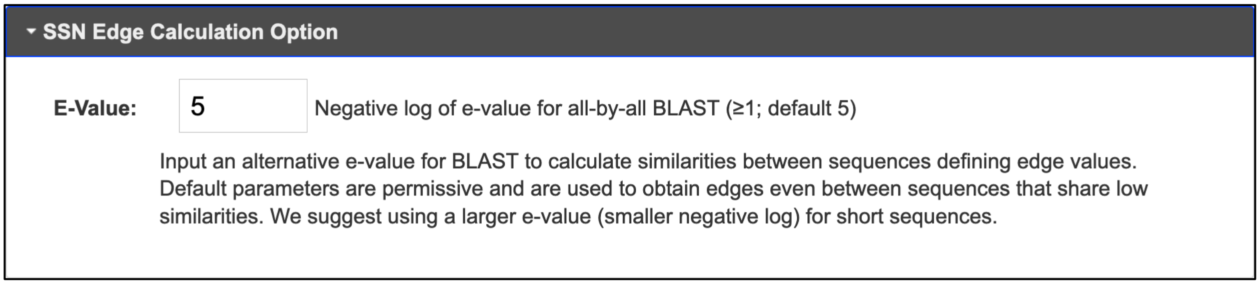


▾ Protein Family Option

Fraction: 1 ? Reduce the number of sequences used to a fraction of the full family size (≥ 1; default: 1)

Selects every Nth sequence in the family; the sequences are assumed to be added randomly to UniProt, so the selected sequences are assumed to be a representative sampling of the family. This allows reduction of the size of the SSN.

SSN Edge Calculation Option. The user can select an alternative e-value (negative base-10 log) for calculating the edges by entering an integer in the "**E-Value**" box (default is 5). If the sequences in the dataset are short, a value <5 should be entered. A larger value might be selected for calculating the edges if the alignment score to generate the SSN is known—there is no need to calculate an alignment score smaller than the alignment score used to generate the SSN (the negative base-10 log of an e-value and the alignment score are similar in magnitude).



▾ SSN Edge Calculation Option

E-Value: 5 Negative log of e-value for all-by-all BLAST (≥1; default 5)

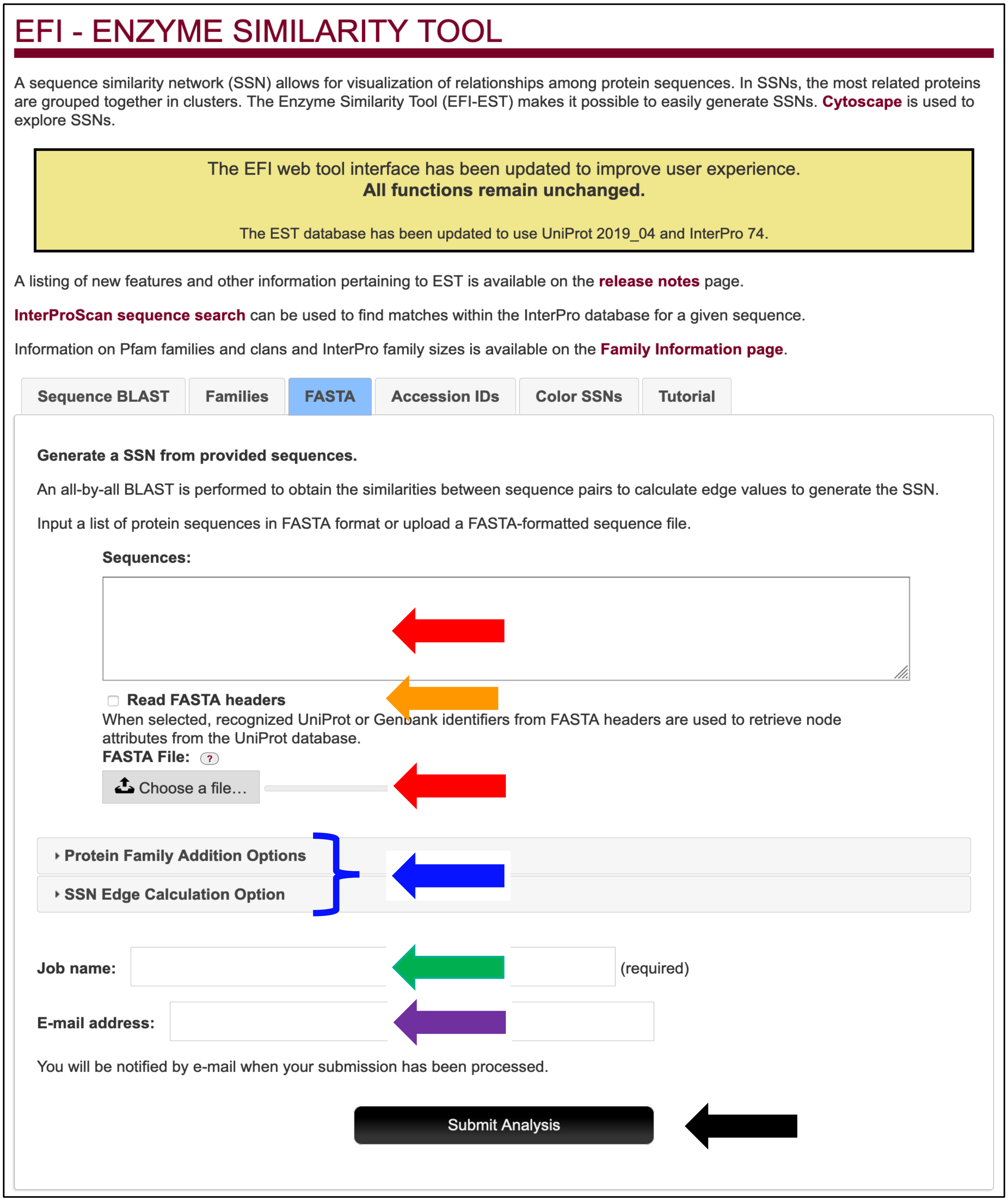
Input an alternative e-value for BLAST to calculate similarities between sequences defining edge values. Default parameters are permissive and are used to obtain edges even between sequences that share low similarities. We suggest using a larger e-value (smaller negative log) for short sequences.

A “**Job name**” (green arrow) is required that will be used on the “**Previous Jobs**” panel on the EFI-EST home page and the “**Job History**” page. On the “**Previous Jobs**” panel and the “**Job History**” page, these jobs are designated “Families”.

An “**E-mail address**” (magenta arrow) is required. The job is initiated with the “**Submit Analysis**” button (black arrow).

The progress of the job is available on the “**Previous Jobs**” panel: “**PENDING**” when the job is received, “**RUNNING**” while the job is running, and a link to the results page when the job is finished. Also, an e-mail is sent to the user when the job is started; a second e-mail is sent when the job is finished that includes a link to the results page.

C. EFI-EST Home Page, “FASTA” tab: User-supplied FASTA file (Option C)



EFI - ENZYME SIMILARITY TOOL

A sequence similarity network (SSN) allows for visualization of relationships among protein sequences. In SSNs, the most related proteins are grouped together in clusters. The Enzyme Similarity Tool (EFI-EST) makes it possible to easily generate SSNs. **Cytoscape** is used to explore SSNs.

The EFI web tool interface has been updated to improve user experience.
All functions remain unchanged.

The EST database has been updated to use UniProt 2019_04 and InterPro 74.

A listing of new features and other information pertaining to EST is available on the **release notes** page.

InterProScan sequence search can be used to find matches within the InterPro database for a given sequence.

Information on Pfam families and clans and InterPro family sizes is available on the **Family Information page**.

Generate a SSN from provided 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 protein sequences in FASTA format or upload a FASTA-formatted sequence file.

Sequences:

□ **Read FASTA headers**
When selected, recognized UniProt or Genbank identifiers from FASTA headers are used to retrieve node attributes from the UniProt database.
FASTA File: ?

Job name: (required)

E-mail address:

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

Option C allows the user to generate an SSN for sequences provided in FASTA format (list or uploaded file).

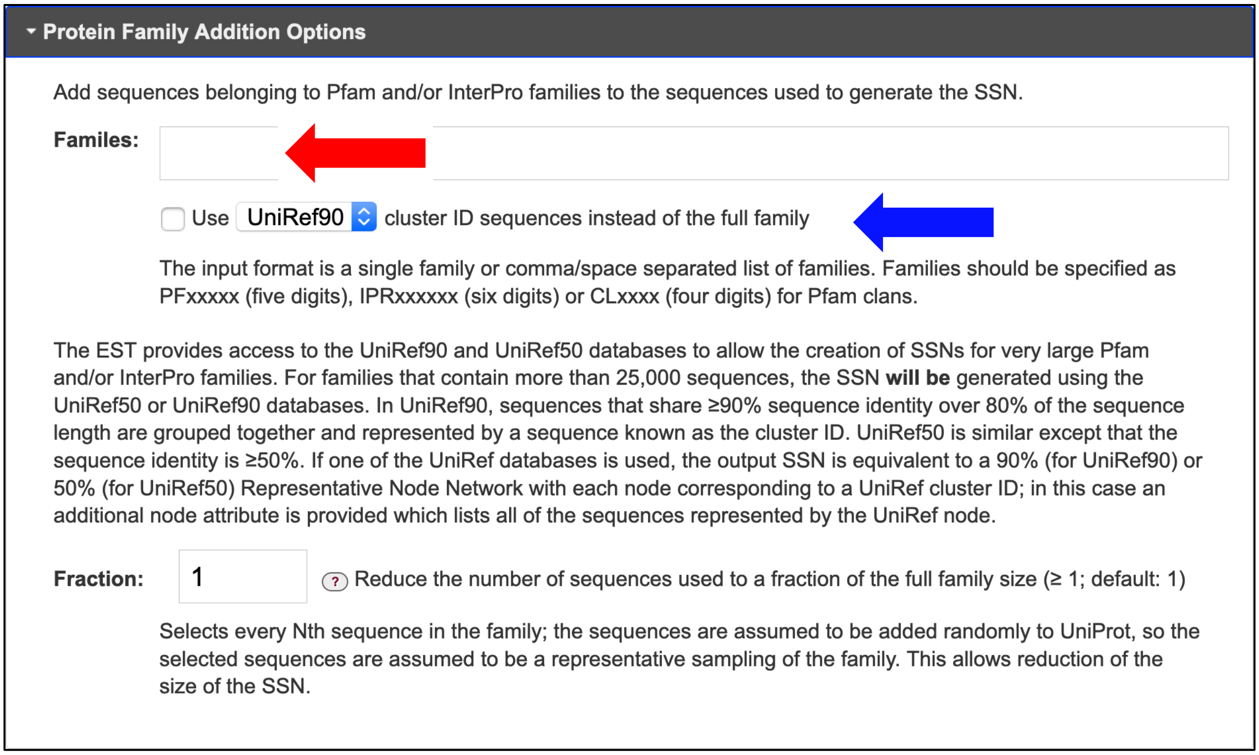
FASTA File Entry. The SSN is generated from user-provided sequences in the FASTA format; these can be pasted in the "**Sequences**" box or uploaded in a text file ("**FASTA File**") (red arrows). FASTA files might be obtained from BLAST analyses using the NCBI database (https://blast.ncbi.nlm.nih.gov/Blast.cgi). By default, irrespective of whether a UniProt ID in the EFI-EST database can be associated with the information in the header (however, see next paragraph), the sequence used for the BLAST is the sequence provided in the FASTA file. The sequences in the FASTA file cannot be associated with node attribute information in the UniProt database (a UniProt ID is required); therefore, only four node attributes are provided in the SSN: "Sequence Source" (identified as "USER"), "Description", the text in the FASTA header; "Sequence Length", the number of residues; and "Sequence", the sequence. In the SSN the nodes are labeled sequentially znnnnnn, with nnnnnn increasing from 000000 in the order of occurrence in the input file; this label is used as the "name" and "shared name" node attributes.

Checking the "**Read FASTA headers**" box (orange arrow) specifies that the FASTA headers be read to identify UniProt or NCBI IDs, if they are present and in an acceptable format (Table S2). If a UniProt or an NCBI ID can be associated with an entry in the UniProt database (using the idmapping file provided by InterPro that identifies equivalent UniProt IDs for NCBI IDs), the node will be labeled with the UniProt ID (instead of znnnnnn; *vide supra*), and the sequence used in the BLAST and the information for the node attributes will be retrieved from the UniProt database. The ID in the header (UniProt or NCBI) will be included in the "Query IDs" node attribute if a UniProt ID is identified; if a UniProt ID cannot be identified, the description in the header will be included in the "Other IDs" node attribute. In both cases, the sequence from the UniProt database will be included in the "Sequence" node attribute. Because the NCBI database contains approximately twice the number of entries in the UniProt database, not all entries with an NCBI ID can be associated with an "equivalent" UniProt accession.

We have noticed that FASTA files uploaded by some users contain edited sequences (not the full-length sequences in UniProt) although the FASTA header contains a valid UniProt ID. If the "**Read FASTA headers**" option is not selected, the EFI-EST scripts use the user-provided sequences to generate the SSN. However, if the "**Read FASTA headers**" option is selected, the scripts retrieve the sequences from the UniProt database for those IDs (UniProt or NCBI) that can be associated with entries in the UniProt database, i.e., *the sequence in the FASTA file is not used*.

The default parameters can be changed using the options in the accordion windows below the box for the input sequence (blue arrow and bracket): "**Protein Family Addition Options**" and "**SSN Edge Calculation Option**".

Protein Family Addition Options. The sequences in the FASTA file can be added to one or more Pfam and/or InterPro family(ies) and/or Pfam clans so that the sequences can be placed in the context of an entire protein family. The family/class ID is entered in the "**Families**" box (red arrow): Pfam family, **PFnnnnn** (five numbers); InterPro family, **IPRnnnnnn** (six numbers); Pfam clan, **CLnnnn** (four numbers).



▾ Protein Family Addition Options

Add sequences belonging to Pfam and/or InterPro families to the sequences used to generate the SSN.

Familes:

□ Use UniRef90 cluster ID sequences instead of the full family

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.

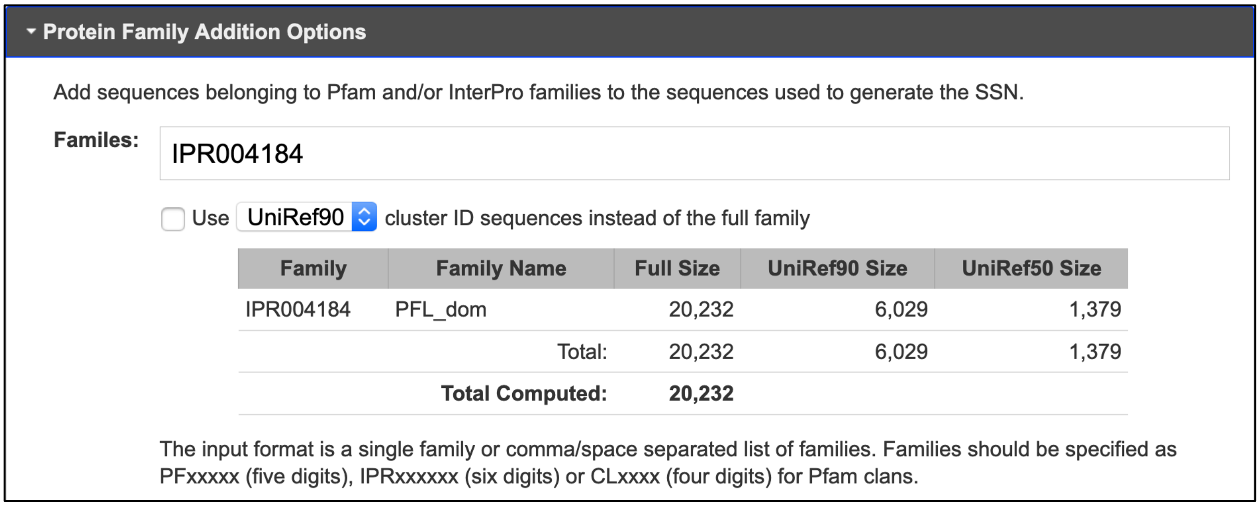
The EST provides access to the UniRef90 and UniRef50 databases to allow the creation of SSNs for very large Pfam and/or InterPro families. For families that contain more than 25,000 sequences, the SSN **will be** generated using the UniRef50 or UniRef90 databases. In UniRef90, sequences that share ≥90% sequence identity over 80% of the sequence length are grouped together and represented by a sequence known as the cluster ID. UniRef50 is similar except that the sequence identity is ≥50%. If one of the UniRef databases is used, the output SSN is equivalent to a 90% (for UniRef90) or 50% (for UniRef50) Representative Node Network with each node corresponding to a UniRef cluster ID; in this case an additional node attribute is provided which lists all of the sequences represented by the UniRef node.

Fraction: 1 ? Reduce the number of sequences used to a fraction of the full family size (≥ 1; default: 1)

Selects every Nth sequence in the family; the sequences are assumed to be added randomly to UniProt, so the selected sequences are assumed to be a representative sampling of the family. This allows reduction of the size of the SSN.

Depending on the number of sequences in the family (*vide infra* for restrictions on the total number of sequences in the input dataset), UniProt IDs, UniRef90 cluster IDs, or UniRef50 cluster IDs can be selected using the check box and UniRef90/UniRef50 menu (blue arrow in the "**Protein Family Additions Option**" accordion).

For example, addition of a UniProt family (IPR004184, the glycyl radical enzyme superfamily example used in the main text):



▾ Protein Family Addition Options

Add sequences belonging to Pfam and/or InterPro families to the sequences used to generate the SSN.

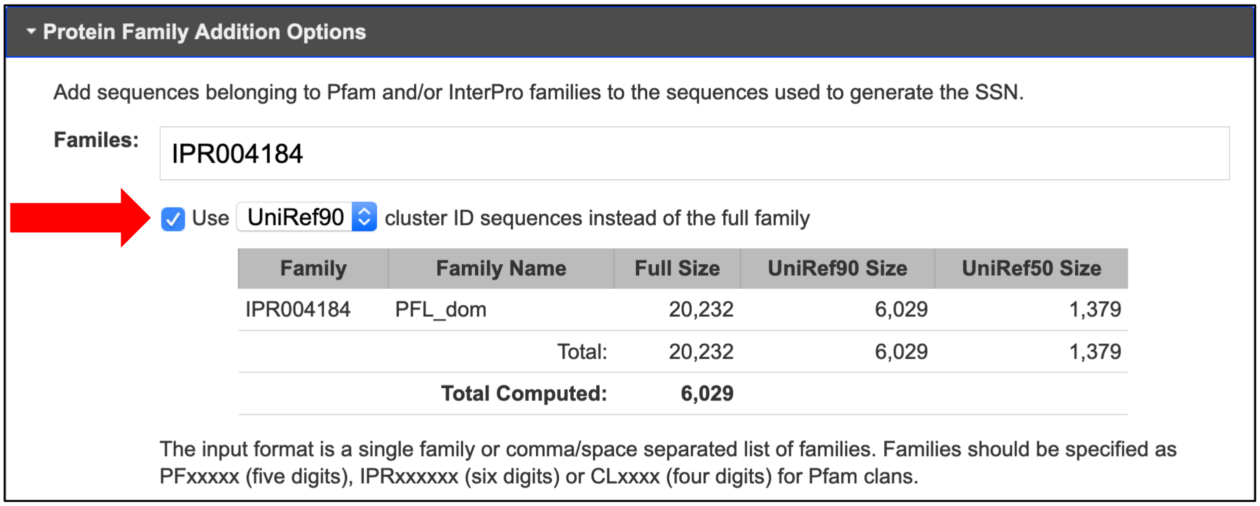
Familes: IPR004184

□ Use UniRef90 cluster ID sequences instead of the full family

Family	Family Name	Full Size	UniRef90 Size	UniRef50 Size
IPR004184	PFL_dom	20,232	6,029	1,379
	Total:	20,232	6,029	1,379
	Total Computed:	**20,232**		

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.

If either UniRef90 or UniRef50 cluster IDs are selected/imposed (red arrow), the user must confirm the selection of these before the job is submitted (blue arrow).



▾ Protein Family Addition Options

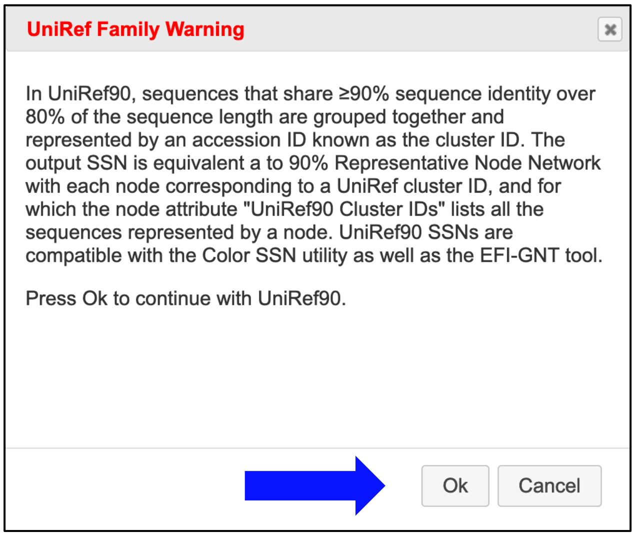
Add sequences belonging to Pfam and/or InterPro families to the sequences used to generate the SSN.

Familes: IPR004184

☑ Use UniRef90 cluster ID sequences instead of the full family

Family	Family Name	Full Size	UniRef90 Size	UniRef50 Size
IPR004184	PFL_dom	20,232	6,029	1,379
	Total:	20,232	6,029	1,379
	Total Computed:	**6,029**		

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.

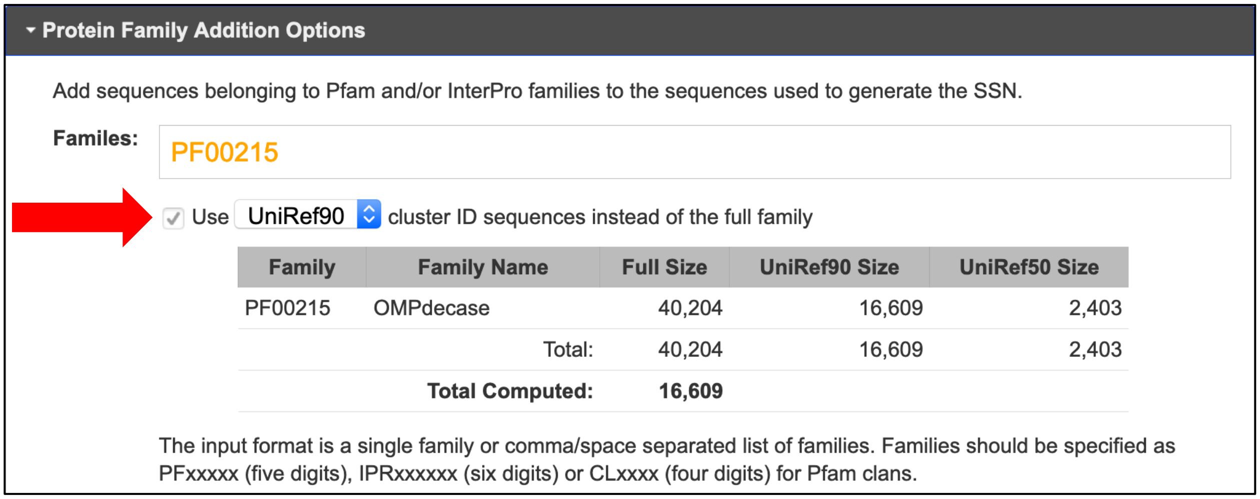


UniRef Family Warning

In UniRef90, sequences that share ≥90% sequence identity over 80% of the sequence length are grouped together and represented by an accession ID known as the cluster ID. The output SSN is equivalent a to 90% Representative Node Network with each node corresponding to a UniRef cluster ID, and for which the node attribute "UniRef90 Cluster IDs" lists all the sequences represented by a node. UniRef90 SSNs are compatible with the Color SSN utility as well as the EFI-GNT tool.

Press Ok to continue with UniRef90.

For a family with >25,000 sequences (red arrow), UniRef90 cluster IDs are required and will be imposed if the number of UniRef90 clusters is ≤100,000. The user must confirm the selection before the job is submitted (blue arrow).



Protein Family Addition Options

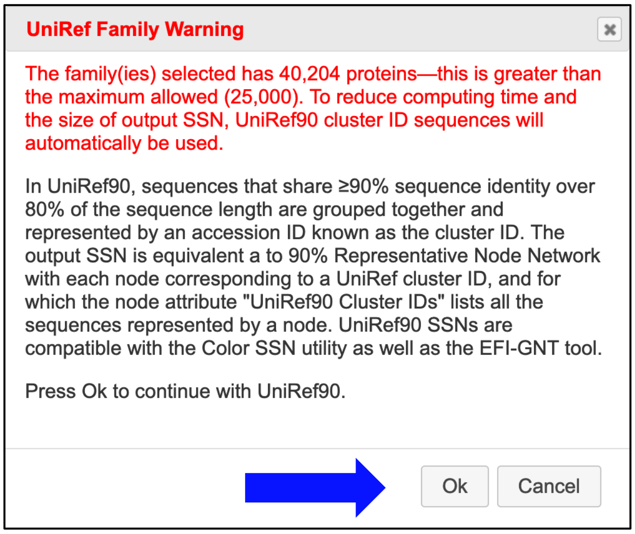
Add sequences belonging to Pfam and/or InterPro families to the sequences used to generate the SSN.

Familes: PF00215

☑ Use UniRef90 cluster ID sequences instead of the full family

Family	Family Name	Full Size	UniRef90 Size	UniRef50 Size
PF00215	OMPdecase	40,204	16,609	2,403
	Total:	40,204	16,609	2,403
	Total Computed:	**16,609**		

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.



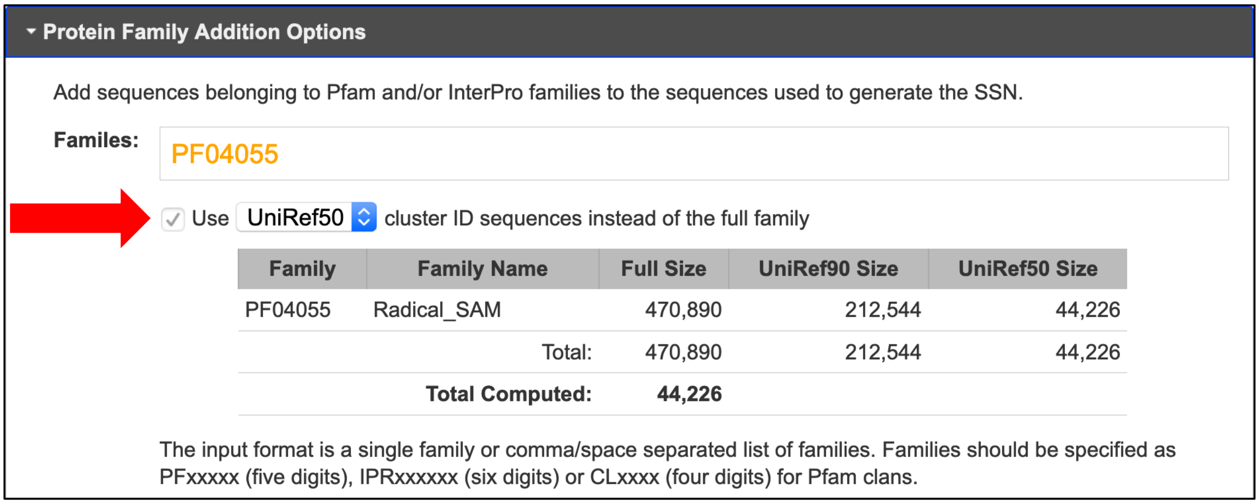
UniRef Family Warning

The family(ies) selected has 40,204 proteins—this is greater than the maximum allowed (25,000). To reduce computing time and the size of output SSN, UniRef90 cluster ID sequences will automatically be used.

In UniRef90, sequences that share ≥90% sequence identity over 80% of the sequence length are grouped together and represented by an accession ID known as the cluster ID. The output SSN is equivalent a to 90% Representative Node Network with each node corresponding to a UniRef cluster ID, and for which the node attribute "UniRef90 Cluster IDs" lists all the sequences represented by a node. UniRef90 SSNs are compatible with the Color SSN utility as well as the EFI-GNT tool.

Press Ok to continue with UniRef90.

For families with >100,000 UniRef90 cluster IDs (red arrow), UniRef50 cluster IDs are required and will be imposed if the number of UniRef50 clusters is ≤100,000. The user must confirm the selection before the job is submitted (blue arrow).



▾ Protein Family Addition Options

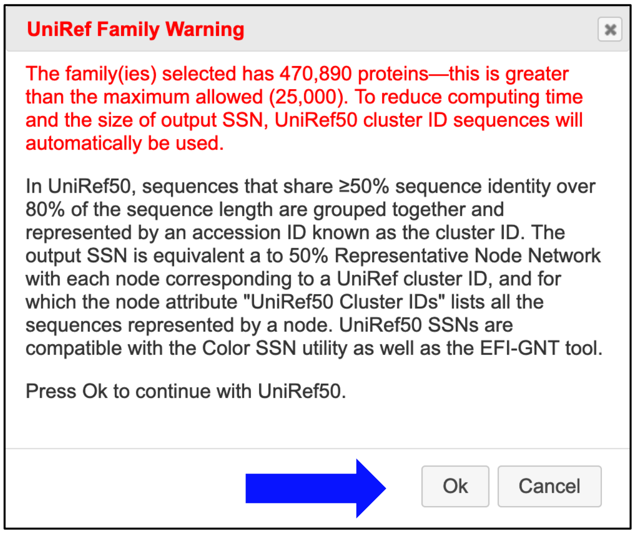
Add sequences belonging to Pfam and/or InterPro families to the sequences used to generate the SSN.

Familes: PF04055

☑ Use UniRef50 cluster ID sequences instead of the full family

Family	Family Name	Full Size	UniRef90 Size	UniRef50 Size
PF04055	Radical_SAM	470,890	212,544	44,226
	Total:	470,890	212,544	44,226
	Total Computed:	**44,226**		

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.



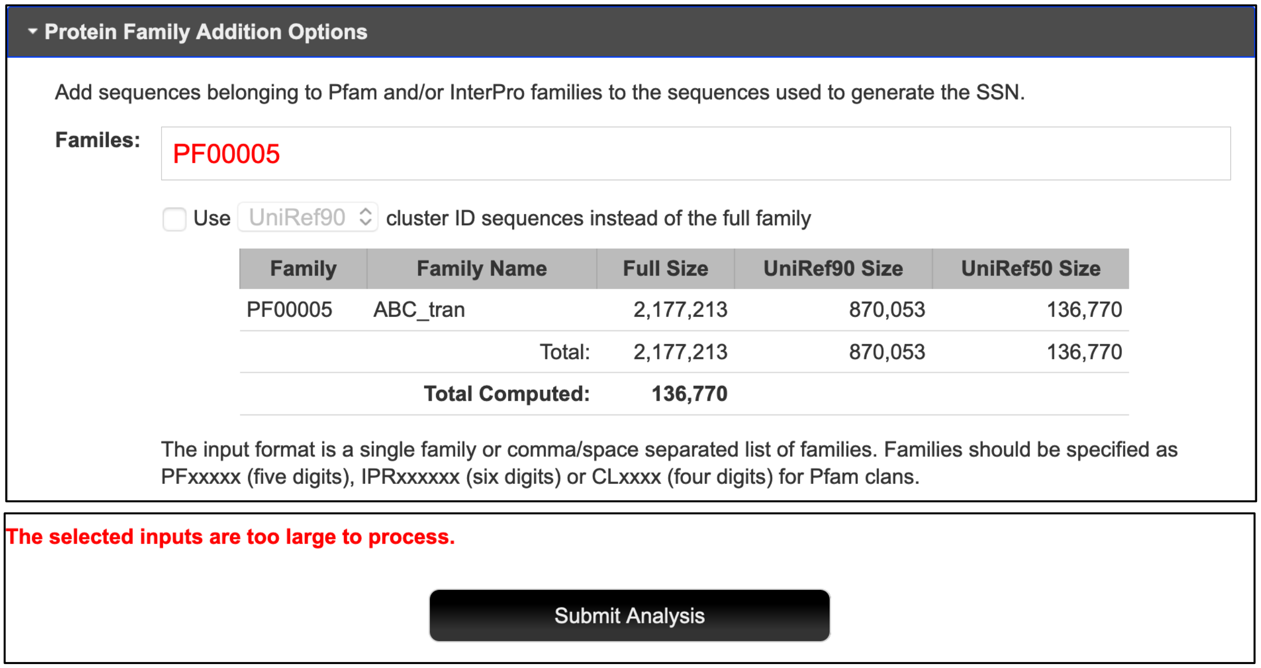
UniRef Family Warning

The family(ies) selected has 470,890 proteins—this is greater than the maximum allowed (25,000). To reduce computing time and the size of output SSN, UniRef50 cluster ID sequences will automatically be used.

In UniRef50, sequences that share ≥50% sequence identity over 80% of the sequence length are grouped together and represented by an accession ID known as the cluster ID. The output SSN is equivalent a to 50% Representative Node Network with each node corresponding to a UniRef cluster ID, and for which the node attribute "UniRef50 Cluster IDs" lists all the sequences represented by a node. UniRef50 SSNs are compatible with the Color SSN utility as well as the EFI-GNT tool.

Press Ok to continue with UniRef50.

A maximum of 100,000 UniRef50 clusters is imposed—the job cannot be submitted when the "Submit Analysis" button is clicked.



▾ Protein Family Addition Options

Add sequences belonging to Pfam and/or InterPro families to the sequences used to generate the SSN.

Familes: PF00005

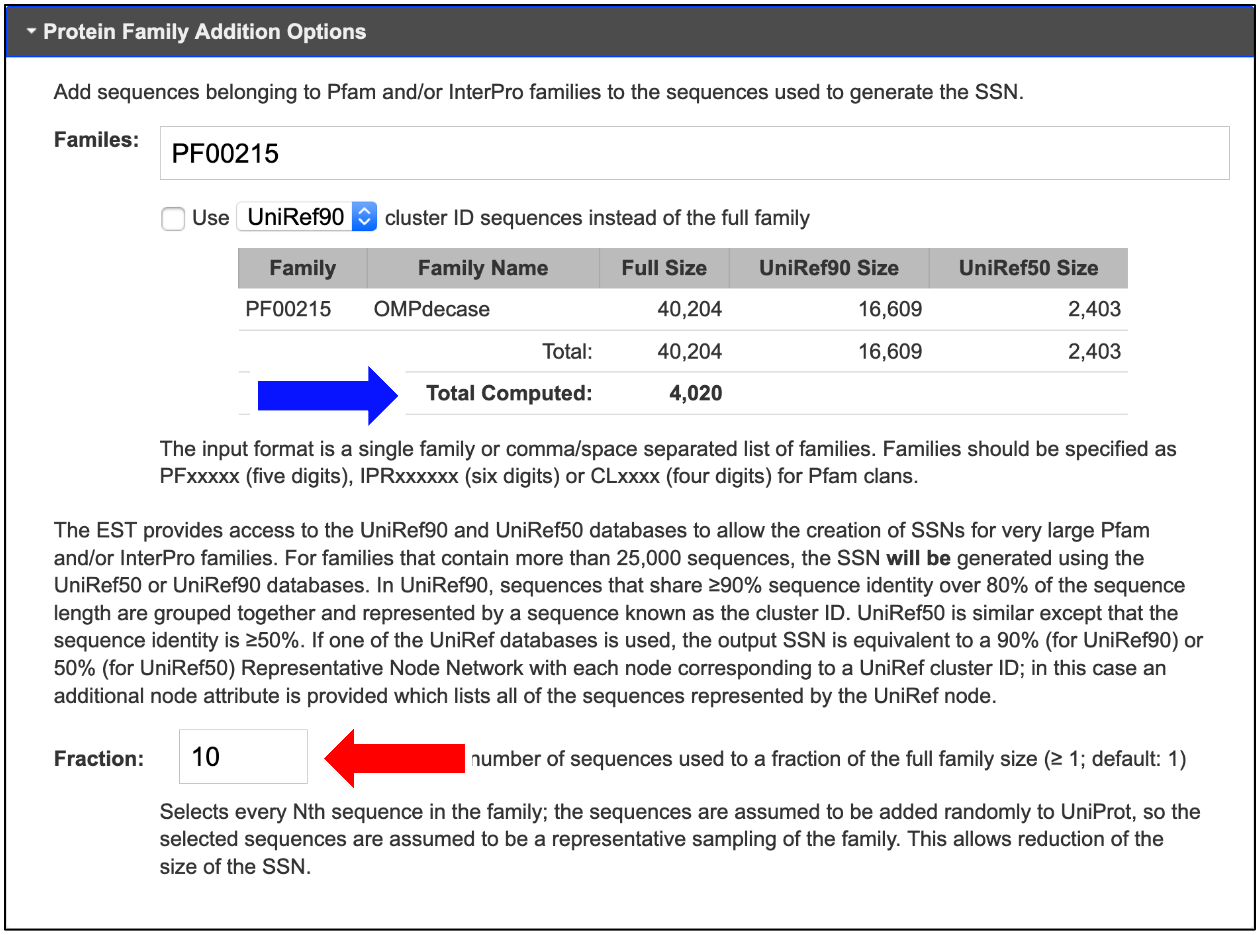
□ Use UniRef90 cluster ID sequences instead of the full family

Family	Family Name	Full Size	UniRef90 Size	UniRef50 Size
PF00005	ABC_tran	2,177,213	870,053	136,770
	Total:	2,177,213	870,053	136,770
	Total Computed:	**136,770**		

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.

The selected inputs are too large to process.

The user also can specify that a representative fraction of the UniProt IDs in the Pfam and/or InterPro family(ies) be added to the BLAST hits: an integer (N) is entered in the "**Fraction**" box (red arrow), with EFI-EST selecting every Nth sequence in the family; the sequences are assumed to be added randomly to UniProt, so the selected sequences are assumed to be a representative sampling of the family (blue arrow). Using a family fraction allows the BLAST hits to be placed in the context of a large family. When using the Fraction option, SwissProt-curated proteins are always included. *The fraction option and UniRef90 or UniRef50 cluster IDs cannot be used together*.



▾ **Protein Family Addition Options**

Add sequences belonging to Pfam and/or InterPro families to the sequences used to generate the SSN.

Familes: PF00215

□ Use UniRef90 cluster ID sequences instead of the full family

Family	Family Name	Full Size	UniRef90 Size	UniRef50 Size
PF00215	OMPdecase	40,204	16,609	2,403
	Total:	40,204	16,609	2,403
	Total Computed:	**4,020**		

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.

The EST provides access to the UniRef90 and UniRef50 databases to allow the creation of SSNs for very large Pfam and/or InterPro families. For families that contain more than 25,000 sequences, the SSN **will be** generated using the UniRef50 or UniRef90 databases. In UniRef90, sequences that share ≥90% sequence identity over 80% of the sequence length are grouped together and represented by a sequence known as the cluster ID. UniRef50 is similar except that the sequence identity is ≥50%. If one of the UniRef databases is used, the output SSN is equivalent to a 90% (for UniRef90) or 50% (for UniRef50) Representative Node Network with each node corresponding to a UniRef cluster ID; in this case an additional node attribute is provided which lists all of the sequences represented by the UniRef node.

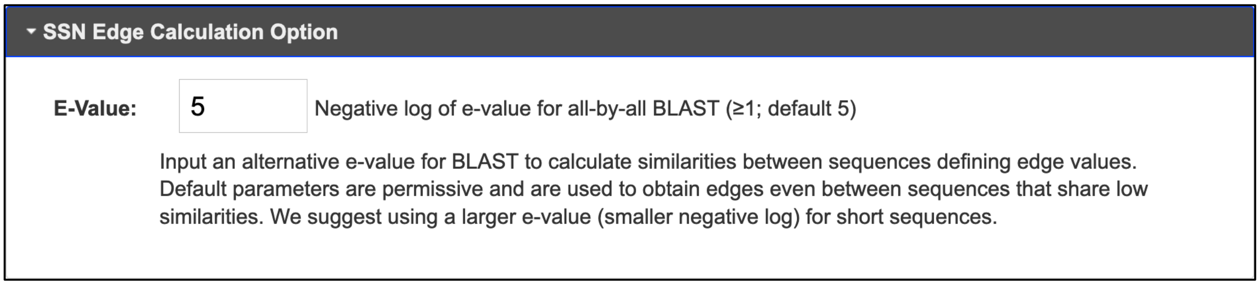
Fraction: 10 number of sequences used to a fraction of the full family size (≥ 1; default: 1)

Selects every Nth sequence in the family; the sequences are assumed to be added randomly to UniProt, so the selected sequences are assumed to be a representative sampling of the family. This allows reduction of the size of the SSN.

In the SSN, the “Sequence Source” node attribute “FAMILY” identifies the sequences from the added Pfam and/or InterPro family(ies). If a family is added using UniProt IDs and the UniProt IDs for a “USER” sequence and a “FAMILY” sequence are identical, the full SSN will contain a merged node with “FAMILY+USER” as the “Sequence Source”; the “User IDs in Cluster” node attribute also will identify the “USER” ID. For a rep node SSN, the “Sequence Source” node attribute for the metanode will be “FAMILY+USER” if it contains one or more merged “FAMILY+USER” nodes; the “User IDs in Cluster” node attribute will be a list of “USER” IDs in the metanode. (In the rep node SSNs, the “name”/”shared” name of the metanode is the ID of the longest sequence in the metanode.)

If a family is added using UniRef clusters and the accession ID for a “USER” sequence is the same as either the UniRef cluster ID or a UniProt ID contained in the UniRef cluster, the full SSN will contain a single merged node with “FAMILY+USER” identifying the “Sequence Source”; the “User IDs in Cluster” node attribute will provide a list of the “USER” IDs in the node. For a rep node SSN, the “Sequence Source” node attribute for the metanode will be “FAMILY+USER” if it contains one or more merged “FAMILY+USER” nodes; the “User IDs in Cluster” node attribute will be a list of “USER” IDs in the metanode. (In the rep node SSNs, the “name”/”shared” name of the metanode is the ID of the longest cluster ID in the metanode.)

SSN Edge Calculation Option. The user can select an alternative e-value (negative base-10 log) for calculating the edges by entering an integer in the “**E-Value**” box (default is 5). If the sequences in the dataset are short, a value <5 should be entered. A larger value might be selected for calculating the edges if the alignment score to generate the SSN is known—there is no need to calculate an alignment score smaller than the alignment score used to generate the SSN (the negative base-10 log of an e-value and the alignment score are similar in magnitude).



E-Value: 5 Negative log of e-value for all-by-all BLAST (≥1; default 5)

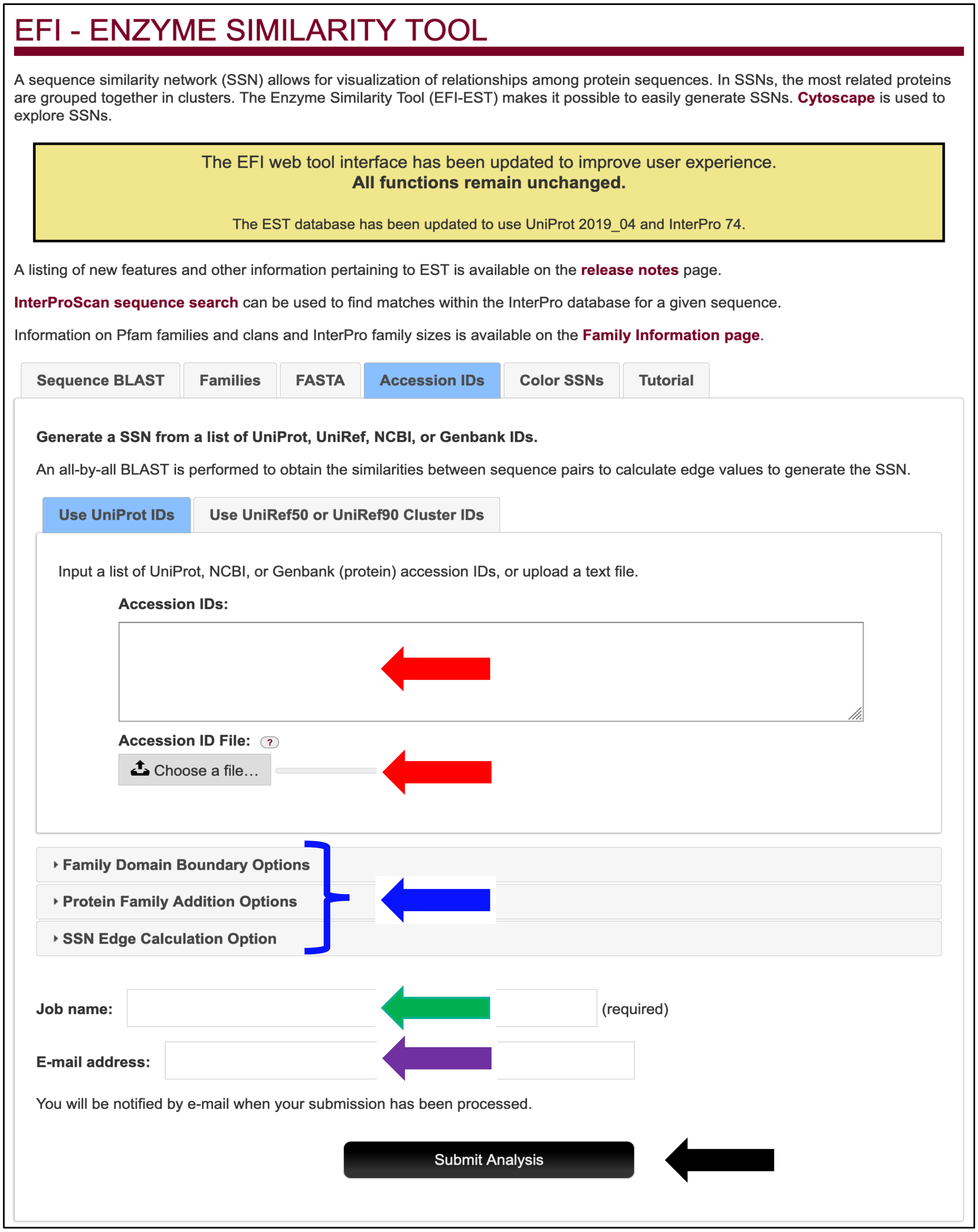
Input an alternative e-value for BLAST to calculate similarities between sequences defining edge values. Default parameters are permissive and are used to obtain edges even between sequences that share low similarities. We suggest using a larger e-value (smaller negative log) for short sequences.

A “**Job name**” (green arrow) is required that will be used on the “**Previous Jobs**” panel on the EFI-EST home page and the “**Job History**” page. On the “**Previous Jobs**” panel and the “**Job History**” page, these jobs are designated “FASTA” if the FASTA header is not interrogated for accession IDs; these jobs are designated “FASTA+Headers” if the FASTA header is interrogated for accession IDs.

An “**E-mail address**” (magenta arrow) is required. The job is initiated with the “**Submit Analysis**” button (black arrow).

The progress of the job is available on the “**Previous Jobs**” panel: “**PENDING**” when the job is received, “**RUNNING**” while the job is running, and a link to the results page when the job is finished. Also, an e-mail is sent to the user when the job is started; a second e-mail is sent when the job is finished that includes a link to the results page.

D. EFI-EST Home Page, “UniProt IDs” tab: List of UniProt and/or NCBI Accession IDs (Option D)



EFI - ENZYME SIMILARITY TOOL

A sequence similarity network (SSN) allows for visualization of relationships among protein sequences. In SSNs, the most related proteins are grouped together in clusters. The Enzyme Similarity Tool (EFI-EST) makes it possible to easily generate SSNs. **Cytoscape** is used to explore SSNs.

The EFI web tool interface has been updated to improve user experience.
All functions remain unchanged.

The EST database has been updated to use UniProt 2019_04 and InterPro 74.

A listing of new features and other information pertaining to EST is available on the **release notes** page.

InterProScan sequence search can be used to find matches within the InterPro database for a given sequence.

Information on Pfam families and clans and InterPro family sizes is available on the **Family Information page**.

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: ?

Choose a file…
Family Domain Boundary Options
Protein Family Addition Options
SSN Edge Calculation Option

Job name: (required)

E-mail address:

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

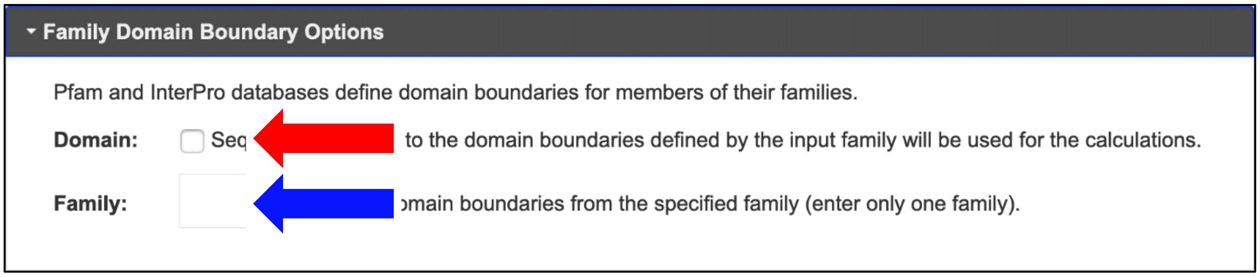
Option D allows the user to generate an SSN from a list of UniProt and/or NCBI IDs (list or uploaded file).

Accession ID Entry. The SSN is generated from a user-provided list of UniProt IDs and/or NCBI IDs; these can be pasted in the "**Accession IDs**" box or uploaded in a text file ("**Accession ID File**") (red arrows). UniProt IDs are used to retrieve sequences and node attribute information from the UniProt database. When recognized via the UniProt-provided "idmapping file", an NCBI ID is used to retrieve the "equivalent" UniProt ID, sequence, and node attribute information. Sequences associated with (NCBI) IDs that cannot be recognized are not included in the SSN; a "nomatch" file listing these IDs is available for download.

Why do we provide Option D? EFI-GNT provides lists of the accession IDs for genome neighbors that are not associated with a Pfam family, so Option D can be used to identify previously uncurated protein families. Also, higher resolution SSNs for clusters in large SSNs can be generated using Option D, allowing access to the node attributes associated with specific UniProt IDs in each cluster in representative node (rep node) SSNs or SSNs generated with UniRef90 or UniRef50 clusters. Colored SSNs generated with the Color SSNs utility and EFI-GNT provide folders of files with lists of the UniProt IDs in each cluster of SSNs generated with UniProt IDs (both full and rep node SSNs), UniProt and UniRef90 IDs in each cluster of SSNs generated using UniRef90 clusters and UniProt, UniRef90, and UniRef50 IDs in each cluster of SSNs generated using UniRef50 clusters (*vide infra*). These files can be used as the input for Option D to generate the higher resolution SSNs.

The default parameters can be changed using the options in the accordion windows below the box for the input sequence (blue arrow and bracket): "**Family Domain Boundaries Options**", "**Protein Family Addition Options**", and "**SSN Edge Calculation Option**".

Family Domain Boundaries Options. By checking the box (red arrow), Option D generates the SSN using sequences of the domains defined by the single family specified in the “**Family**” box [Pfam family, **PFnnnnn** (five numbers); InterPro family, **IPRnnnnnn** (six numbers); or Pfam clan, **CLnnnn** (four numbers)] (blue arrow). An earlier section described both the advantages and disadvantages associated with generating SSNs from domains. W*hen domains are used, the “**Family**” must be a single Pfam family or an InterPro family/domain that is defined by one family database*. In the SSN, the “name” and “shared name” node attributes include the UniProt ID followed by colon-delimited residue numbers for N- and C-termini of the domain: UniProtID:N-terminus:C-terminus.



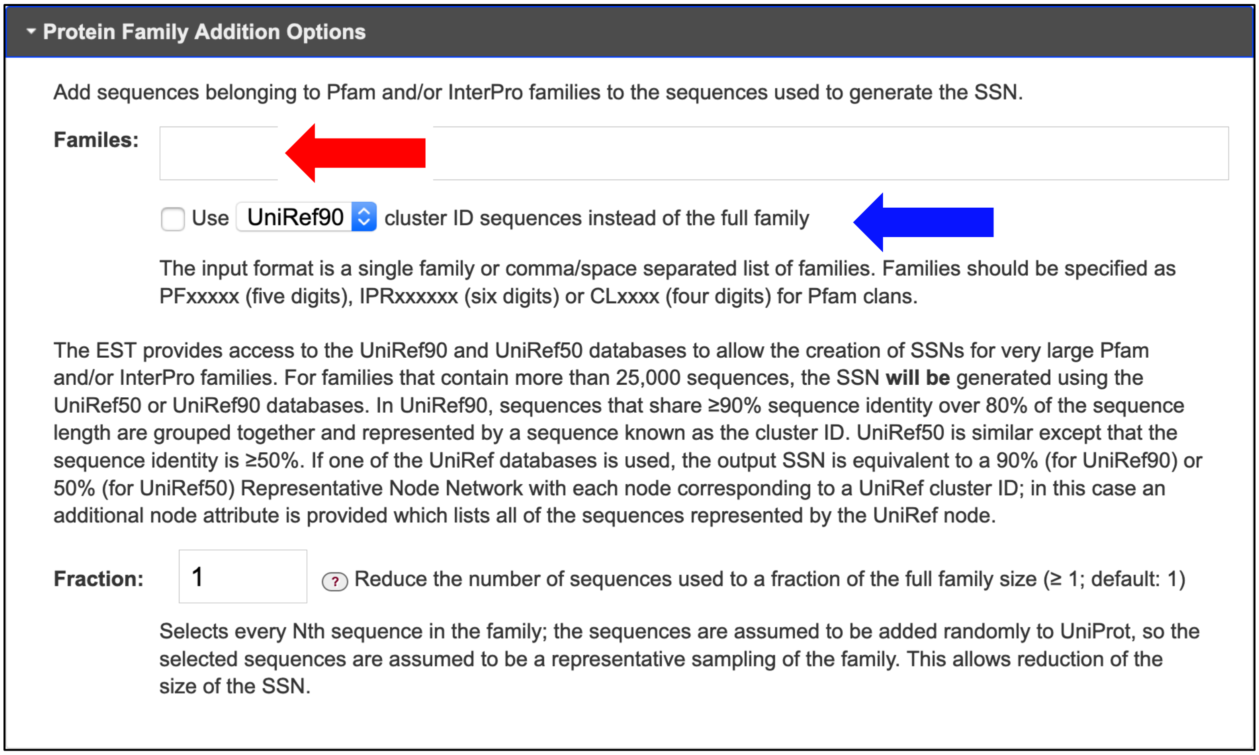
▾ **Family Domain Boundary Options**

Pfam and InterPro databases define domain boundaries for members of their families.

Domain: □ Seq to the domain boundaries defined by the input family will be used for the calculations.

Family: omain boundaries from the specified family (enter only one family).

Protein Family Addition Options. The sequences in the FASTA file can be added to one or more Pfam and/or InterPro family(ies) and/or Pfam clans so that the sequences can be placed in the context of an entire protein family. The family/class ID is entered in the "**Families**" box (red arrow): Pfam family, **PFnnnnn** (five numbers); InterPro family, **IPRnnnnnn** (six numbers); Pfam clan, **CLnnnn** (four numbers).



Add sequences belonging to Pfam and/or InterPro families to the sequences used to generate the SSN.

Familes:

□ Use UniRef90 cluster ID sequences instead of the full family

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.

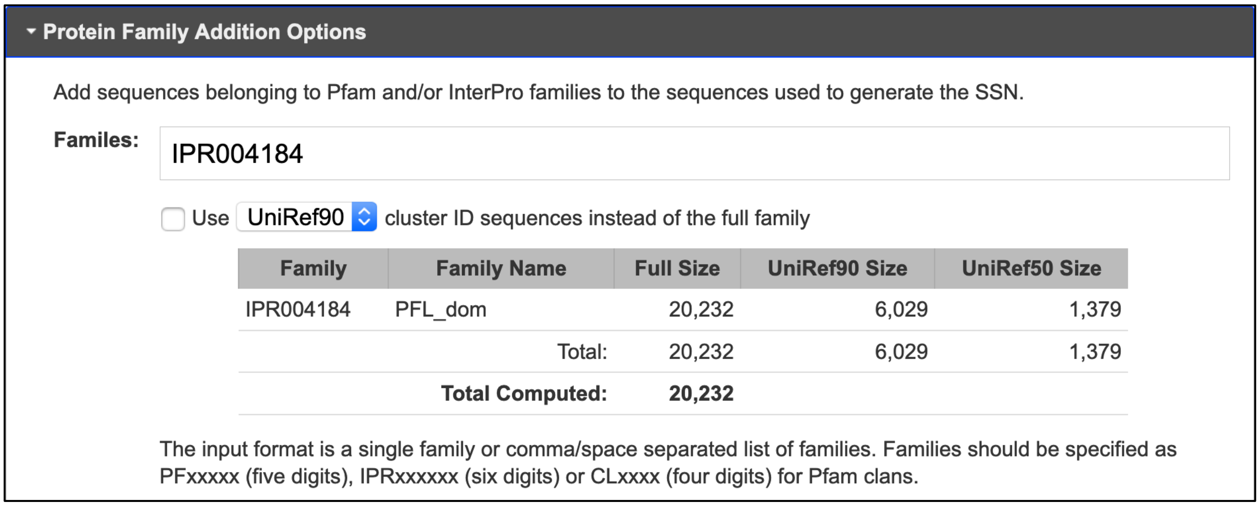
The EST provides access to the UniRef90 and UniRef50 databases to allow the creation of SSNs for very large Pfam and/or InterPro families. For families that contain more than 25,000 sequences, the SSN **will be** generated using the UniRef50 or UniRef90 databases. In UniRef90, sequences that share ≥90% sequence identity over 80% of the sequence length are grouped together and represented by a sequence known as the cluster ID. UniRef50 is similar except that the sequence identity is ≥50%. If one of the UniRef databases is used, the output SSN is equivalent to a 90% (for UniRef90) or 50% (for UniRef50) Representative Node Network with each node corresponding to a UniRef cluster ID; in this case an additional node attribute is provided which lists all of the sequences represented by the UniRef node.

Fraction: 1 ? Reduce the number of sequences used to a fraction of the full family size (≥ 1; default: 1)

Selects every Nth sequence in the family; the sequences are assumed to be added randomly to UniProt, so the selected sequences are assumed to be a representative sampling of the family. This allows reduction of the size of the SSN.

Depending on the number of sequences in the family (*vide infra* for restrictions on the total number of sequences in the input dataset), UniProt IDs, UniRef90 cluster IDs, or UniRef50 cluster IDs can be selected using the check box and UniRef90/UniRef50 menu (blue arrow in the "**Protein Family Additions Option**" accordion).

For example, addition of a UniProt family (IPR004184, the glycyl radical enzyme superfamily example used in the main text):



Add sequences belonging to Pfam and/or InterPro families to the sequences used to generate the SSN.

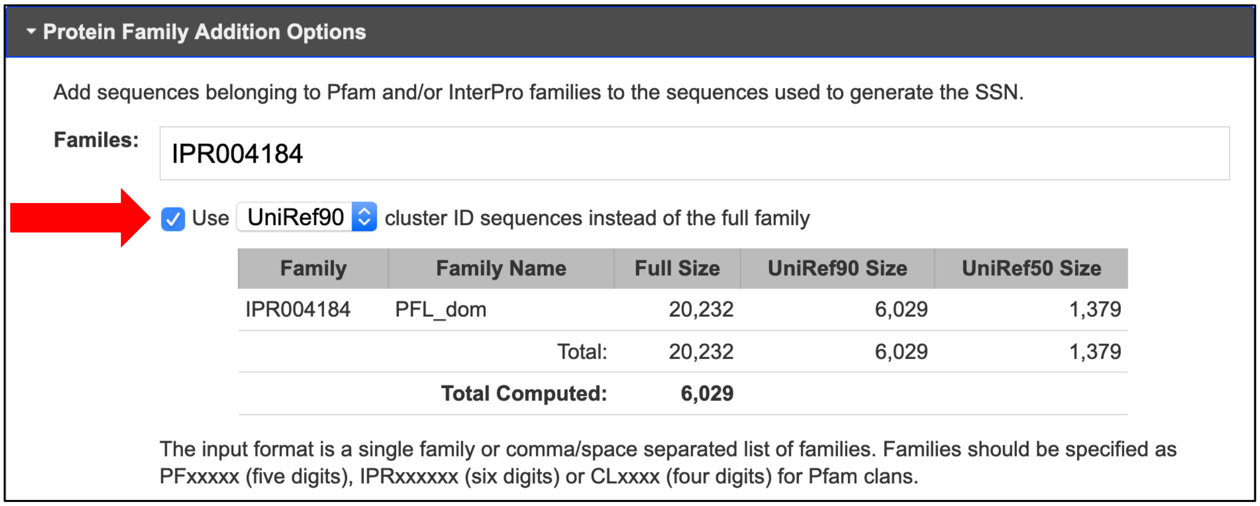
Familes: IPR004184

□ Use UniRef90 cluster ID sequences instead of the full family

Family	Family Name	Full Size	UniRef90 Size	UniRef50 Size
IPR004184	PFL_dom	20,232	6,029	1,379
	Total:	20,232	6,029	1,379
	Total Computed:	**20,232**		

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.

If either UniRef90 or UniRef50 cluster IDs are selected/imposed (red arrow), the user must confirm the selection of these before the job is submitted (blue arrow).



▾ Protein Family Addition Options

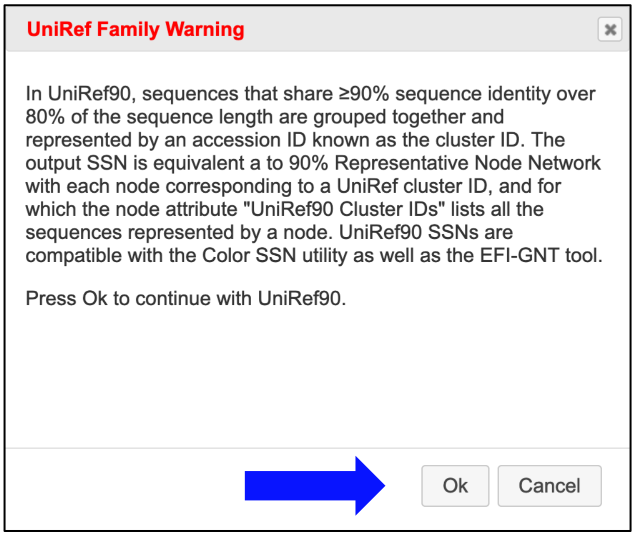
Add sequences belonging to Pfam and/or InterPro families to the sequences used to generate the SSN.

Familes: IPR004184

☑ Use UniRef90 cluster ID sequences instead of the full family

Family	Family Name	Full Size	UniRef90 Size	UniRef50 Size
IPR004184	PFL_dom	20,232	6,029	1,379
	Total:	20,232	6,029	1,379
	Total Computed:	**6,029**		

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.

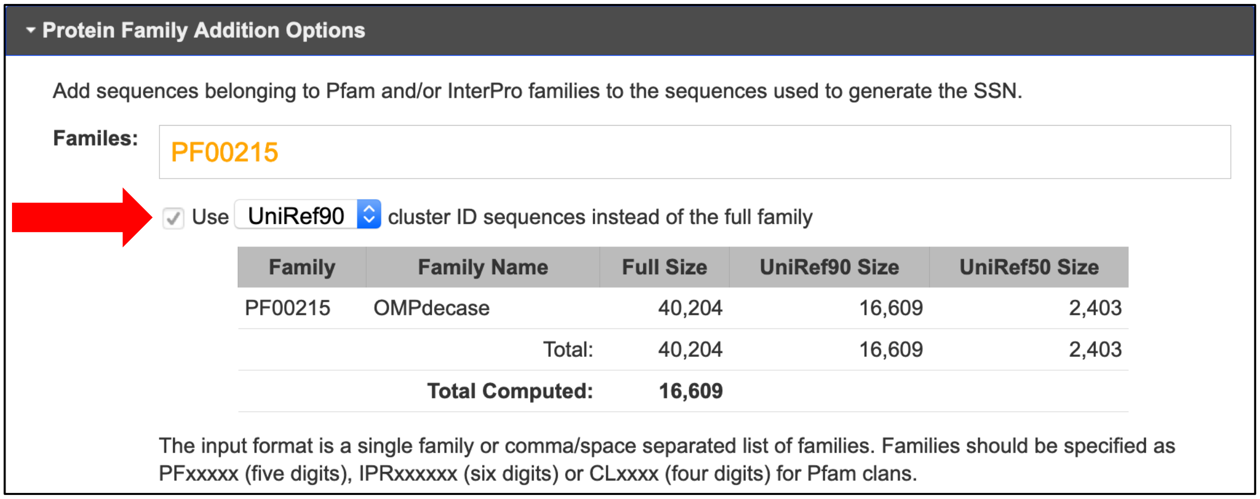


UniRef Family Warning

In UniRef90, sequences that share ≥90% sequence identity over 80% of the sequence length are grouped together and represented by an accession ID known as the cluster ID. The output SSN is equivalent a to 90% Representative Node Network with each node corresponding to a UniRef cluster ID, and for which the node attribute "UniRef90 Cluster IDs" lists all the sequences represented by a node. UniRef90 SSNs are compatible with the Color SSN utility as well as the EFI-GNT tool.

Press Ok to continue with UniRef90.

For a family with >25,000 sequences (red arrow), UniRef90 cluster IDs are required and will be imposed if the number of UniRef90 clusters is ≤100,000. The user must confirm the selection before the job is submitted (blue arrow).



▾ Protein Family Addition Options

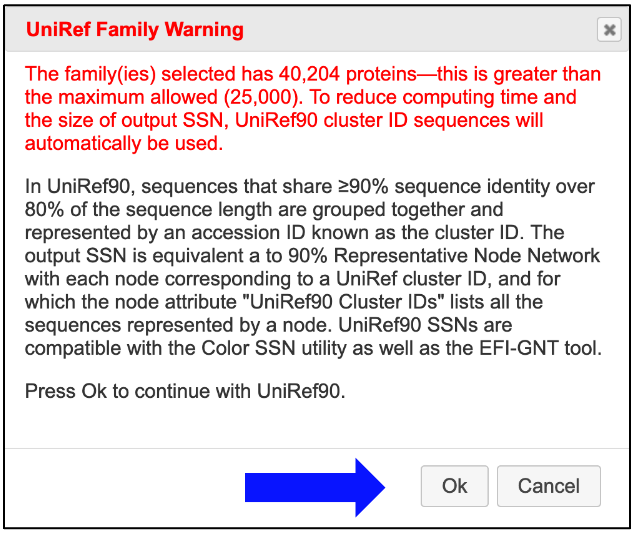
Add sequences belonging to Pfam and/or InterPro families to the sequences used to generate the SSN.

Familes: PF00215

☑ Use UniRef90 cluster ID sequences instead of the full family

Family	Family Name	Full Size	UniRef90 Size	UniRef50 Size
PF00215	OMPdecase	40,204	16,609	2,403
	Total:	40,204	16,609	2,403
	Total Computed:	**16,609**		

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.



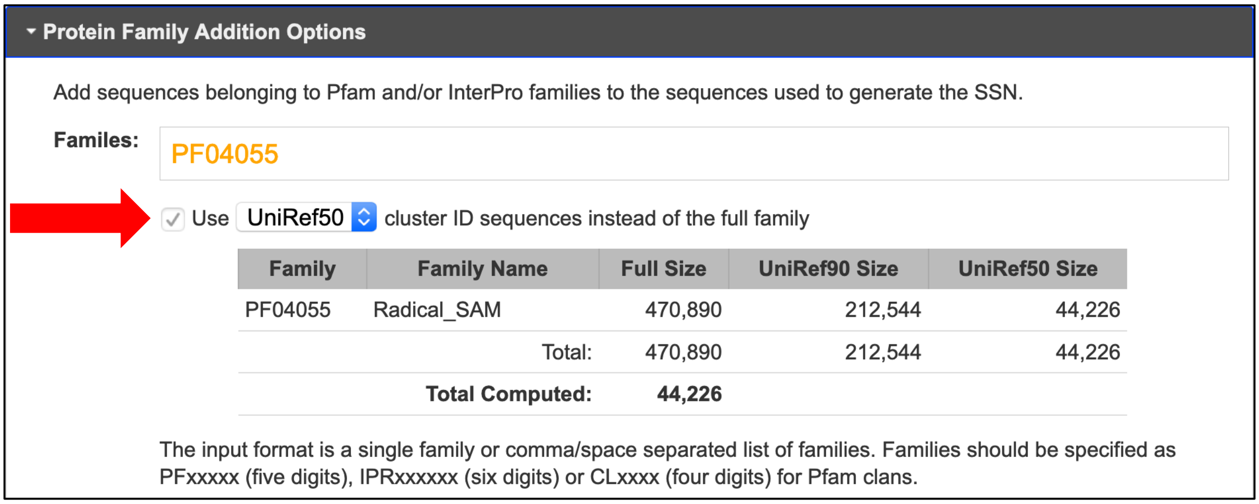
UniRef Family Warning

The family(ies) selected has 40,204 proteins—this is greater than the maximum allowed (25,000). To reduce computing time and the size of output SSN, UniRef90 cluster ID sequences will automatically be used.

In UniRef90, sequences that share ≥90% sequence identity over 80% of the sequence length are grouped together and represented by an accession ID known as the cluster ID. The output SSN is equivalent a to 90% Representative Node Network with each node corresponding to a UniRef cluster ID, and for which the node attribute "UniRef90 Cluster IDs" lists all the sequences represented by a node. UniRef90 SSNs are compatible with the Color SSN utility as well as the EFI-GNT tool.

Press Ok to continue with UniRef90.

For families with >100,000 UniRef90 cluster IDs (red arrow), UniRef50 cluster IDs are required and will be imposed if the number of UniRef50 clusters is ≤100,000. The user must confirm the selection before the job is submitted (blue arrow).



▾ Protein Family Addition Options

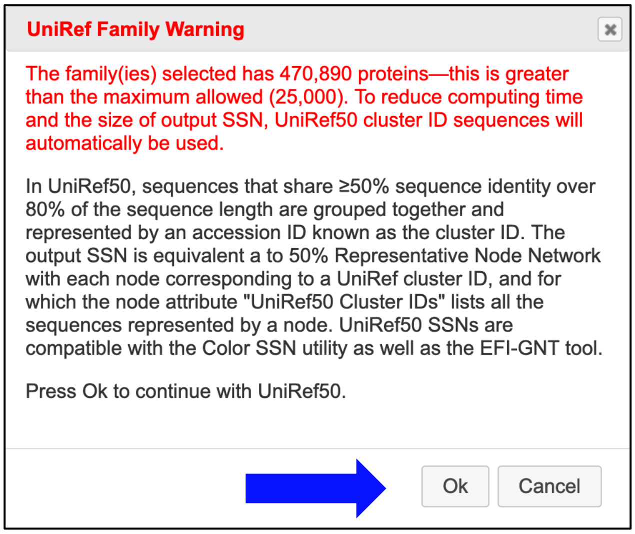
Add sequences belonging to Pfam and/or InterPro families to the sequences used to generate the SSN.

Familes: PF04055

☑ Use UniRef50 cluster ID sequences instead of the full family

Family	Family Name	Full Size	UniRef90 Size	UniRef50 Size
PF04055	Radical_SAM	470,890	212,544	44,226
	Total:	470,890	212,544	44,226
	Total Computed:	**44,226**		

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.



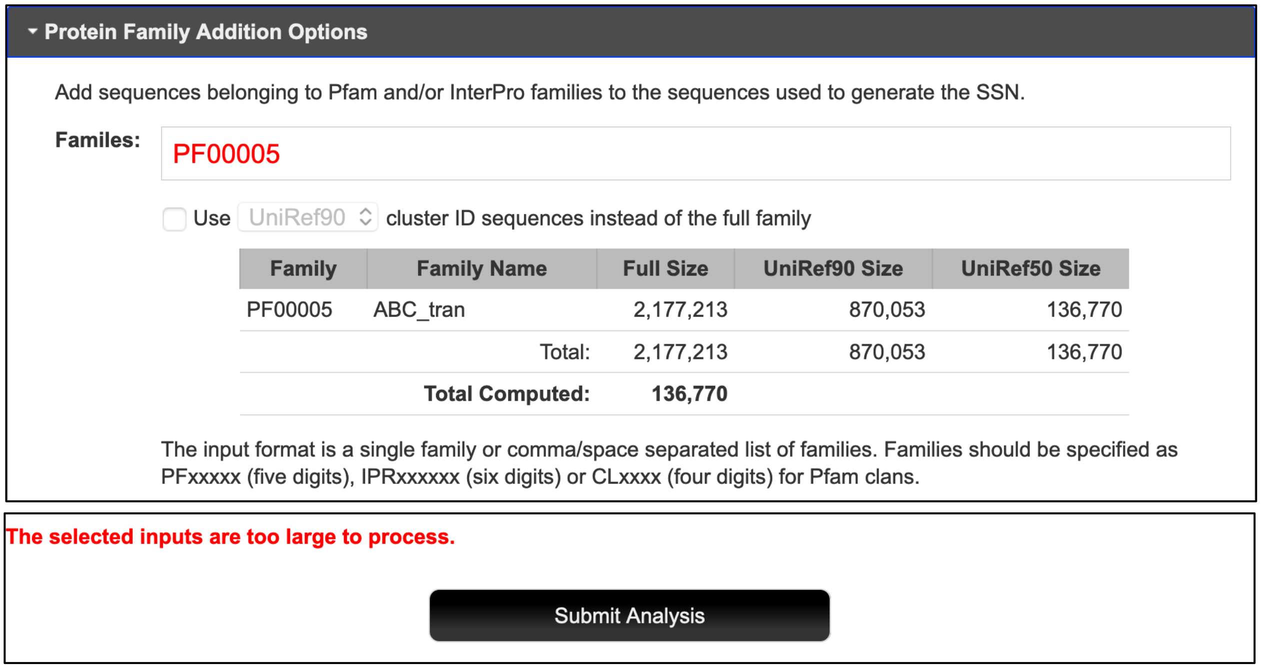
UniRef Family Warning

The family(ies) selected has 470,890 proteins—this is greater than the maximum allowed (25,000). To reduce computing time and the size of output SSN, UniRef50 cluster ID sequences will automatically be used.

In UniRef50, sequences that share ≥50% sequence identity over 80% of the sequence length are grouped together and represented by an accession ID known as the cluster ID. The output SSN is equivalent a to 50% Representative Node Network with each node corresponding to a UniRef cluster ID, and for which the node attribute "UniRef50 Cluster IDs" lists all the sequences represented by a node. UniRef50 SSNs are compatible with the Color SSN utility as well as the EFI-GNT tool.

Press Ok to continue with UniRef50.

A maximum of 100,000 UniRef50 clusters is imposed—the job cannot be submitted when the “Submit Analysis” button is clicked.



▾ Protein Family Addition Options

Add sequences belonging to Pfam and/or InterPro families to the sequences used to generate the SSN.

Familes: PF00005

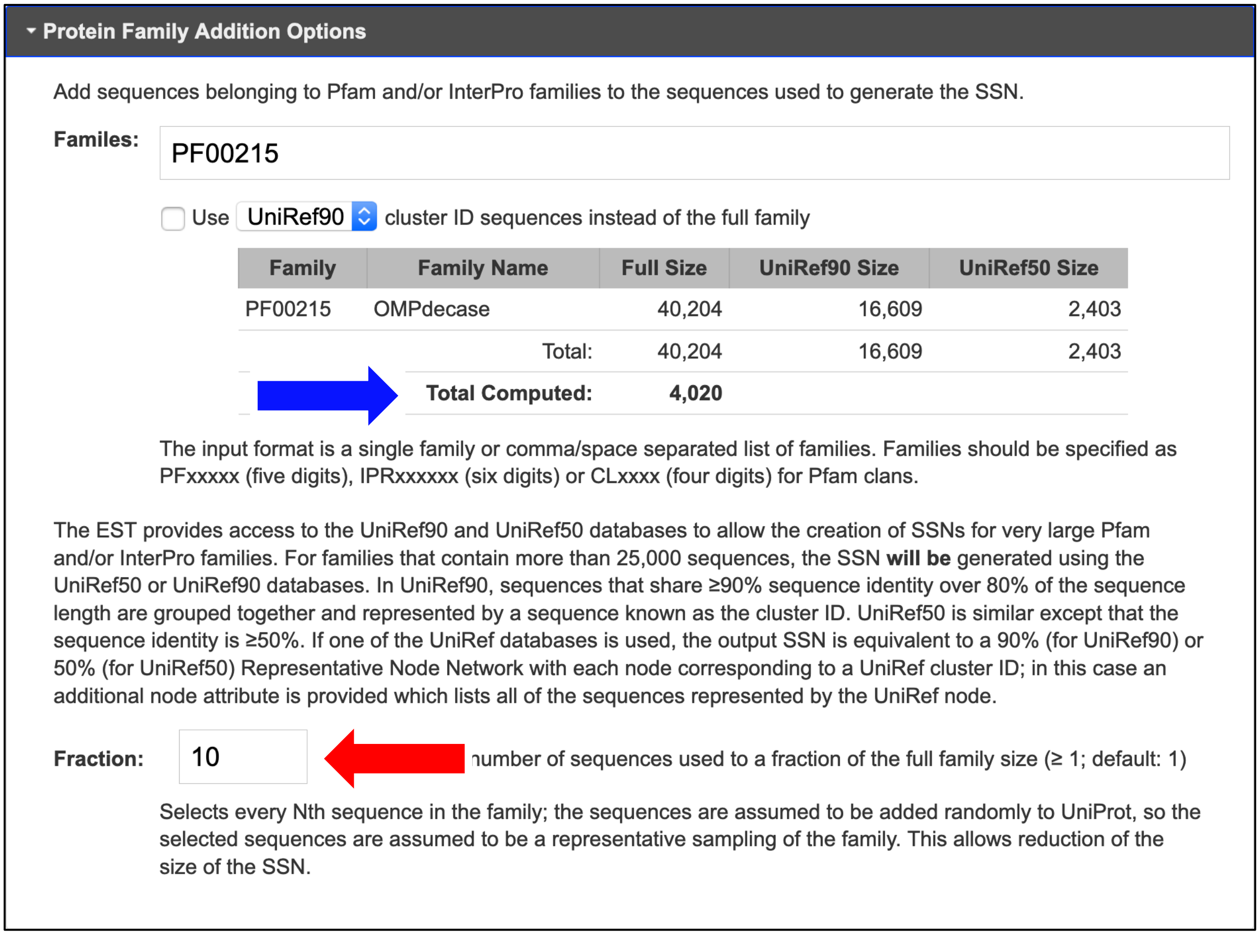
□ Use UniRef90 cluster ID sequences instead of the full family

Family	Family Name	Full Size	UniRef90 Size	UniRef50 Size
PF00005	ABC_tran	2,177,213	870,053	136,770
	Total:	2,177,213	870,053	136,770
	Total Computed:	**136,770**		

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.

The selected inputs are too large to process.

The user also can specify that a representative fraction of the UniProt IDs in the Pfam and/or InterPro family(ies) be added to the BLAST hits: an integer (N) is entered in the "**Fraction**" box (red arrow), with EFI-EST selecting every Nth sequence in the family; the sequences are assumed to be added randomly to UniProt, so the selected sequences are assumed to be a representative sampling of the family (blue arrow). Using a family fraction allows the BLAST hits to be placed in the context of a large family. When using the Fraction option, SwissProt-curated proteins are always included. *The fraction option and UniRef90 or UniRef50 cluster IDs cannot be used together*.



Protein Family Addition Options

Add sequences belonging to Pfam and/or InterPro families to the sequences used to generate the SSN.

Familes: PF00215

□ Use UniRef90 cluster ID sequences instead of the full family

Family	Family Name	Full Size	UniRef90 Size	UniRef50 Size
PF00215	OMPdecase	40,204	16,609	2,403
	Total:	40,204	16,609	2,403
	Total Computed:	**4,020**		

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.

The EST provides access to the UniRef90 and UniRef50 databases to allow the creation of SSNs for very large Pfam and/or InterPro families. For families that contain more than 25,000 sequences, the SSN **will be** generated using the UniRef50 or UniRef90 databases. In UniRef90, sequences that share ≥90% sequence identity over 80% of the sequence length are grouped together and represented by a sequence known as the cluster ID. UniRef50 is similar except that the sequence identity is ≥50%. If one of the UniRef databases is used, the output SSN is equivalent to a 90% (for UniRef90) or 50% (for UniRef50) Representative Node Network with each node corresponding to a UniRef cluster ID; in this case an additional node attribute is provided which lists all of the sequences represented by the UniRef node.

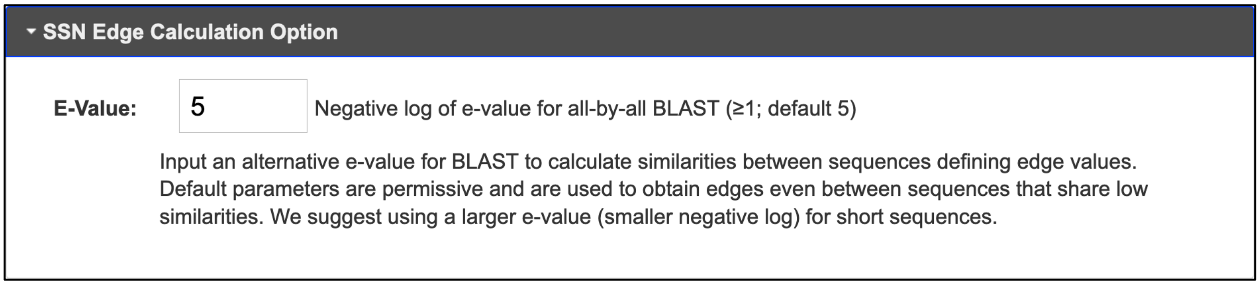
Fraction: 10 number of sequences used to a fraction of the full family size (≥ 1; default: 1)

Selects every Nth sequence in the family; the sequences are assumed to be added randomly to UniProt, so the selected sequences are assumed to be a representative sampling of the family. This allows reduction of the size of the SSN.

In the SSN, the “Sequence Source” node attribute identifies the origin of the sequence: “USER” identifies an input sequence; “FAMILY” identifies the sequences in added Pfam and/or InterPro family(ies). If a family is added using UniProt IDs and the UniProt IDs for a “USER” sequence and a “FAMILY” sequence are identical, the full SSN will contain a merged node with “FAMILY+USER” as the “Sequence Source”; the “User IDs in Cluster” node attribute also will identify the “USER” ID. For a rep node SSN, the “Sequence Source” node attribute for the metanode will be “FAMILY+USER” if it contains one or more merged “FAMILY+USER” nodes; the “User IDs in Cluster” node attribute will be a list of “USER” IDs in the metanode. (In the rep node SSNs, the “name”/”shared” name of the metanode is the ID of the longest sequence in the metanode.)

If a family is added using UniRef clusters and the accession ID for a “USER” sequence is the same as either the UniRef cluster ID or the UniProt ID of a family member contained in the UniRef cluster, the full SSN will contain a single merged node with “FAMILY+USER” identifying the “Sequence Source”; the “User IDs in Cluster” node attribute will provide a list of the “USER” IDs in the node. For a rep node SSN, the “Sequence Source” node attribute for the metanode will be “FAMILY+USER” if it contains one or more merged “FAMILY+USER” nodes; the “User IDs in Cluster” node attribute will be a list of “USER” IDs in the metanode. (In the rep node SSNs, the “name”/”shared” name of the metanode is the ID of the longest cluster ID in the metanode.)

SSN Edge Calculation Option. The user can select an alternative e-value (negative base-10 log) for calculating the edges by entering an integer in the "**E-Value**" box (default is 5). If the sequences in the dataset are short, a value <5 should be entered. A larger value might be selected for calculating the edges if the alignment score to generate the SSN is known—there is no need to calculate an alignment score smaller than the alignment score used to generate the SSN (the negative base-10 log of an e-value and the alignment score are similar in magnitude).



E-Value: 5 Negative log of e-value for all-by-all BLAST (≥1; default 5)

Input an alternative e-value for BLAST to calculate similarities between sequences defining edge values. Default parameters are permissive and are used to obtain edges even between sequences that share low similarities. We suggest using a larger e-value (smaller negative log) for short sequences.

A "**Job name**" (green arrow) is required that will be used on the "**Previous Jobs**" panel on the EFI-EST home page and the "**Job History**" page. On the "**Previous Jobs**" panel and the "**Job History**" page, these jobs are designated "Accession IDs".

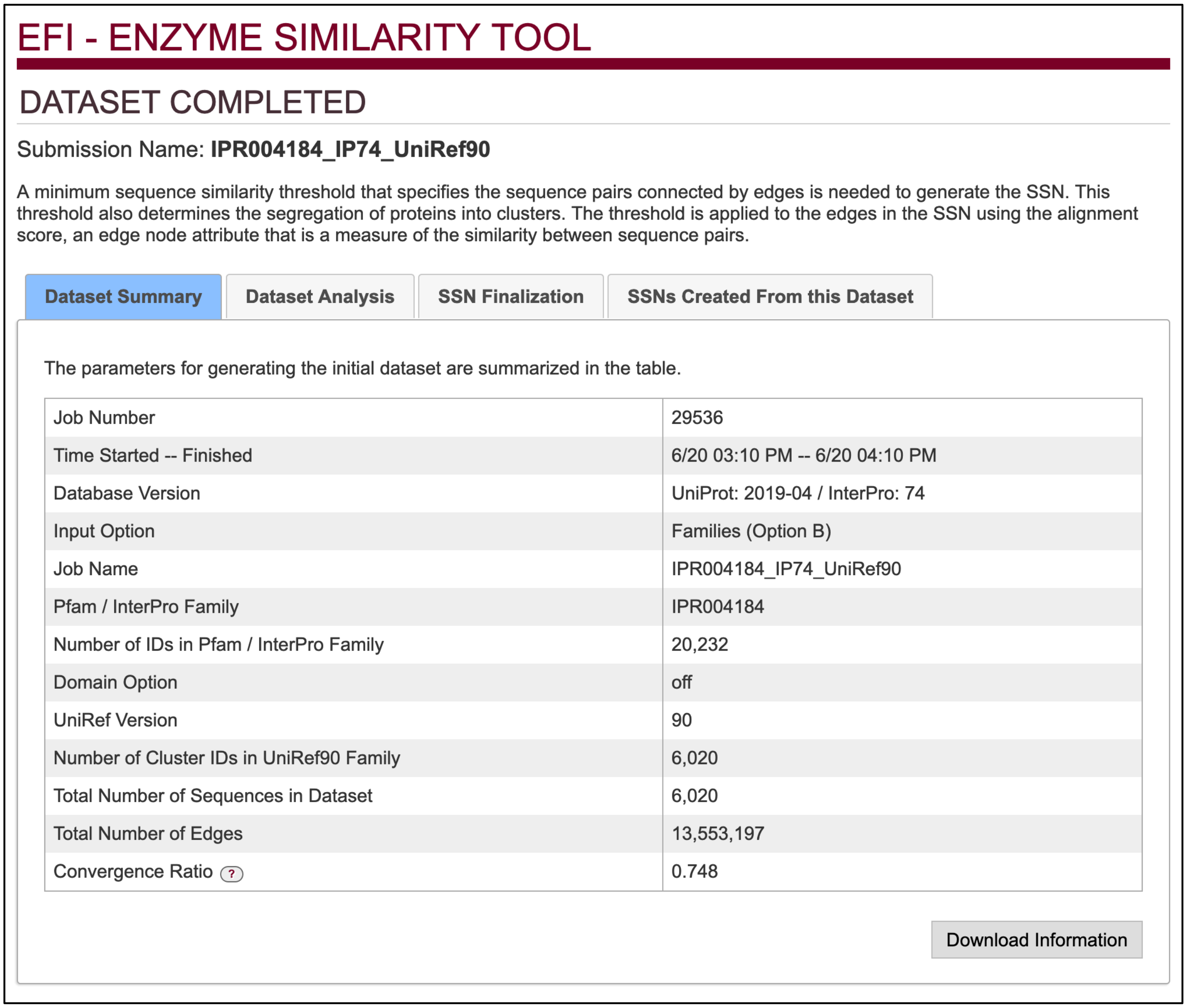
An "**E-mail address**" (magenta arrow) is required. The job is initiated with the "**Submit Analysis**" button (black arrow).

The progress of the job is available on the "**Previous Jobs**" panel: "**PENDING**" when the job is received, "**RUNNING**" while the job is running, and a link to the results page when the job is finished. Also, an e-mail is sent to the user when the job is started; a second e-mail is sent when the job is finished that includes a link to the results page.

E. EFI-EST “Dataset Completed” Page, Selection of minimum/maximum lengths for including sequences and minimum alignment score threshold for drawing edges

When the BLAST is completed, the user is directed to the “**Dataset Completed**” page. This page has two tabs, “**Dataset Summary**” and “**Dataset Analysis**”, that provide information to assist the choice of alignment score and a third tab, “**SSN Finalization**”, to specify the parameters for generating the initial SSN.

“Dataset Summary” tab. This tab provides a table summarizing the identity of the input option (A, B, C, or D), job name, and other parameters specified in generating the input dataset, e.g., Pfam/InterPro family, Fraction option, and/or Domain option; it also provides the number of sequences in the input dataset and the number of edges that were calculated.



EFI - ENZYME SIMILARITY TOOL

DATASET COMPLETED

Submission Name: **IPR004184_IP74_UniRef90**

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.

The parameters for generating the initial dataset are summarized in the table.

Job Number	29536
Time Started -- Finished	6/20 03:10 PM -- 6/20 04:10 PM
Database Version	UniProt: 2019-04 / InterPro: 74
Input Option	Families (Option B)
Job Name	IPR004184_IP74_UniRef90
Pfam / InterPro Family	IPR004184
Number of IDs in Pfam / InterPro Family	20,232
Domain Option	off
UniRef Version	90
Number of Cluster IDs in UniRef90 Family	6,020
Total Number of Sequences in Dataset	6,020
Total Number of Edges	13,553,197
Convergence Ratio ?	0.748

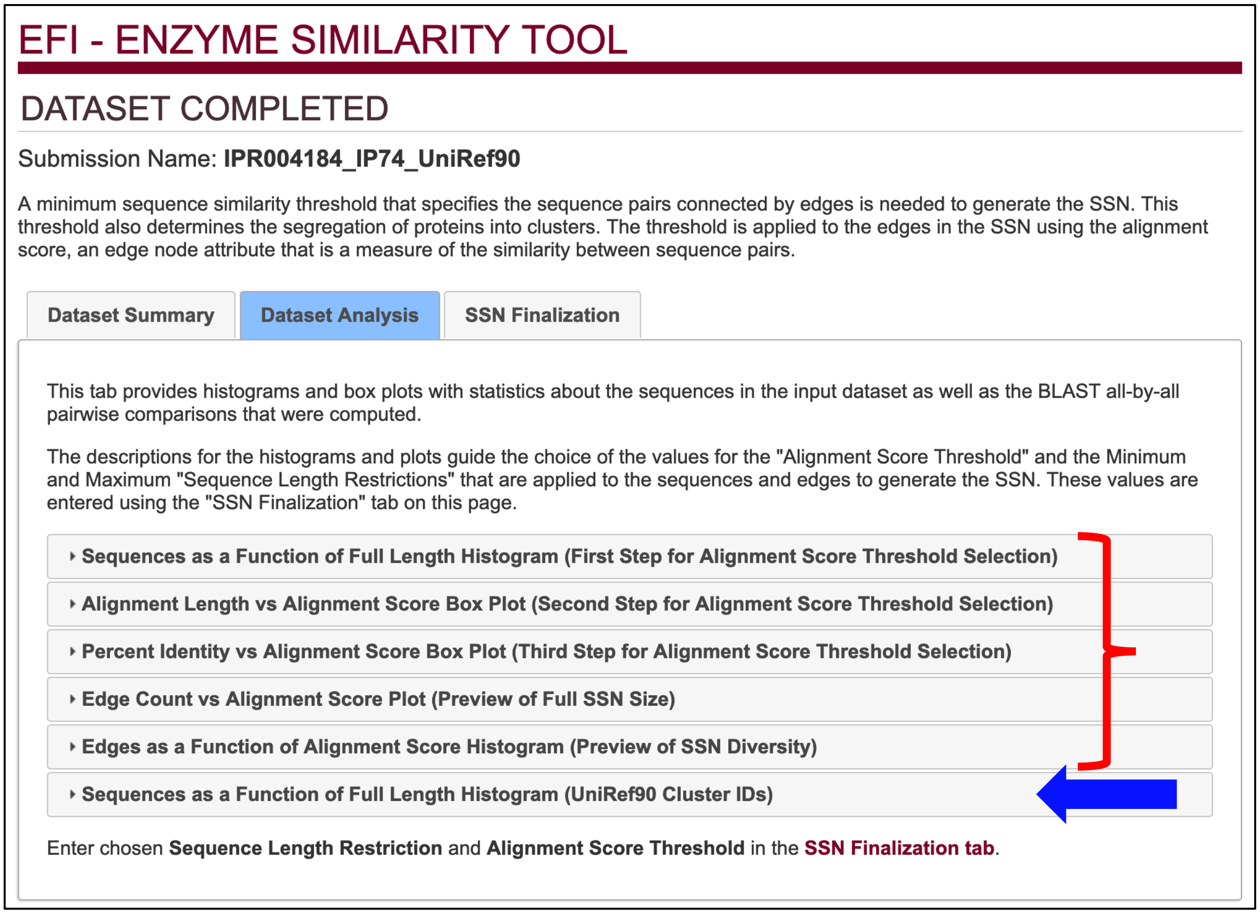
“Dataset Analysis” tab. Irrespective of which Option (A, B, C, or D) was used to generate the input dataset for full-length sequences, the procedure for selecting the alignment score threshold for generating the initial SSN is based on the same considerations: 1) the distribution of sequence lengths in the input dataset, 2) the alignment lengths for the query and subject sequences used by BLAST in the pairwise comparison to generate the edges (should be for full-length proteins), and 3) the relationship between alignment score and pairwise percent identity calculated by BLAST.

Options B and D allow the user to generate SSNs using the domains specified by the input family (one Pfam family or one InterPro family/domain defined by one database; *vide infra*). The procedure for choosing the minimum initial alignment score threshold for domain SSNs is the same as that used for full-length proteins; however, additional histograms are provided to assist the user (Section II.E.2).

Section II.E.1 describes the histograms and plots provided by the "**Dataset Analysis**" tab for SSNs generated with full-length sequences (first section) using Options A, B, C, and D. Section II.E.2 then describes the histograms and plots provided by the "**Dataset Analysis**" tab for SSNs generated with domains using Options B and D.

1. “Dataset Analysis” tab for full-length sequences (Options A, B, C, and D)

Five histograms/boxplots/plots (red bracket) are provided when the full-length sequences of UniProt IDs and/or FASTA files are used to generate the SSN.



EFI - ENZYME SIMILARITY TOOL

DATASET COMPLETED

Submission Name: **IPR004184_IP74_UniRef90**

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.

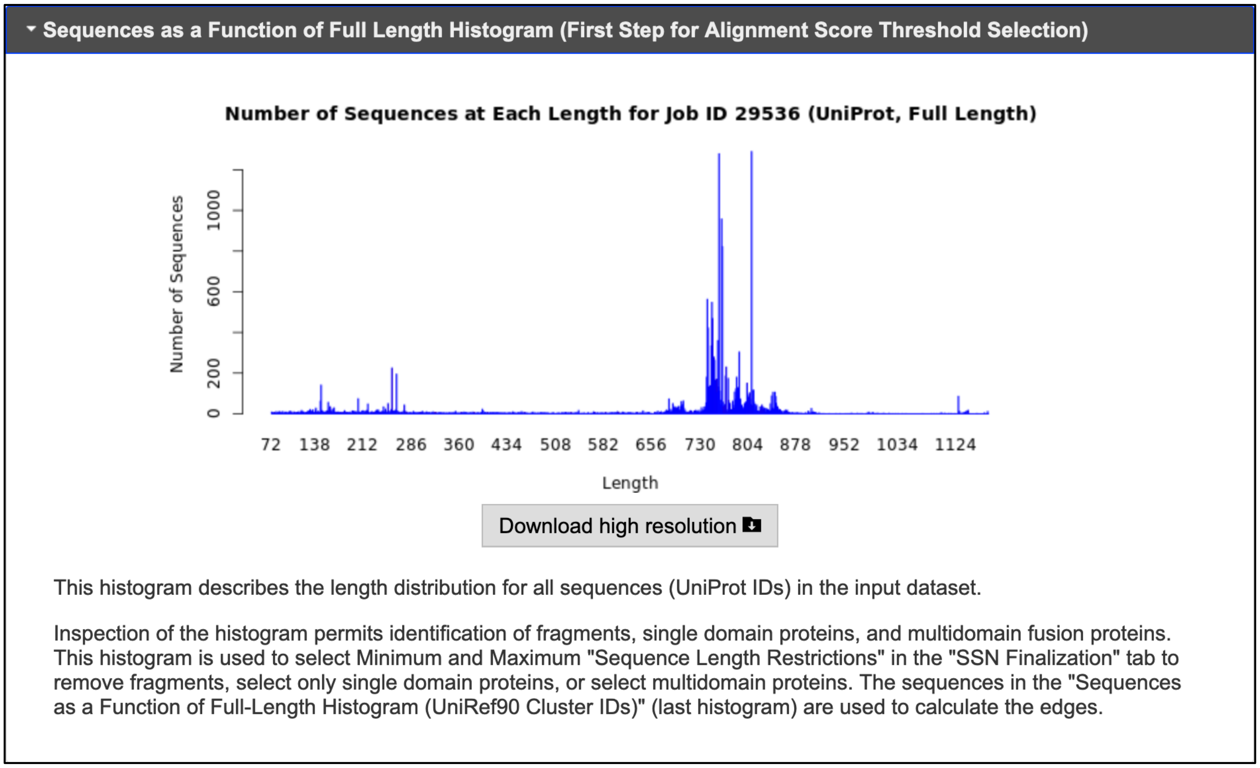
This tab provides histograms and box plots with statistics about the sequences in the input dataset as well as the BLAST all-by-all pairwise comparisons that were computed.

The descriptions for the histograms and plots guide the choice of the values for the "Alignment Score Threshold" and the Minimum and Maximum "Sequence Length Restrictions" that are applied to the sequences and edges to generate the SSN. These values are entered using the "SSN Finalization" tab on this page.

- Sequences as a Function of Full Length Histogram (First Step for Alignment Score Threshold Selection)
- Alignment Length vs Alignment Score Box Plot (Second Step for Alignment Score Threshold Selection)
- Percent Identity vs Alignment Score Box Plot (Third Step for Alignment Score Threshold Selection)
- Edge Count vs Alignment Score Plot (Preview of Full SSN Size)
- Edges as a Function of Alignment Score Histogram (Preview of SSN Diversity)
- Sequences as a Function of Full Length Histogram (UniRef90 Cluster IDs)

Enter chosen **Sequence Length Restriction** and **Alignment Score Threshold** in the **SSN Finalization tab**.

1) “**Sequences as a Function of Full-Length Histogram (First Step for Alignment Score Threshold Selection)**” describes the length distribution of all sequences, e.g., UniProt IDs and/or FASTA sequences, in the input dataset.



▾ Sequences as a Function of Full Length Histogram (First Step for Alignment Score Threshold Selection)

Number of Sequences at Each Length for Job ID 29536 (UniProt, Full Length)

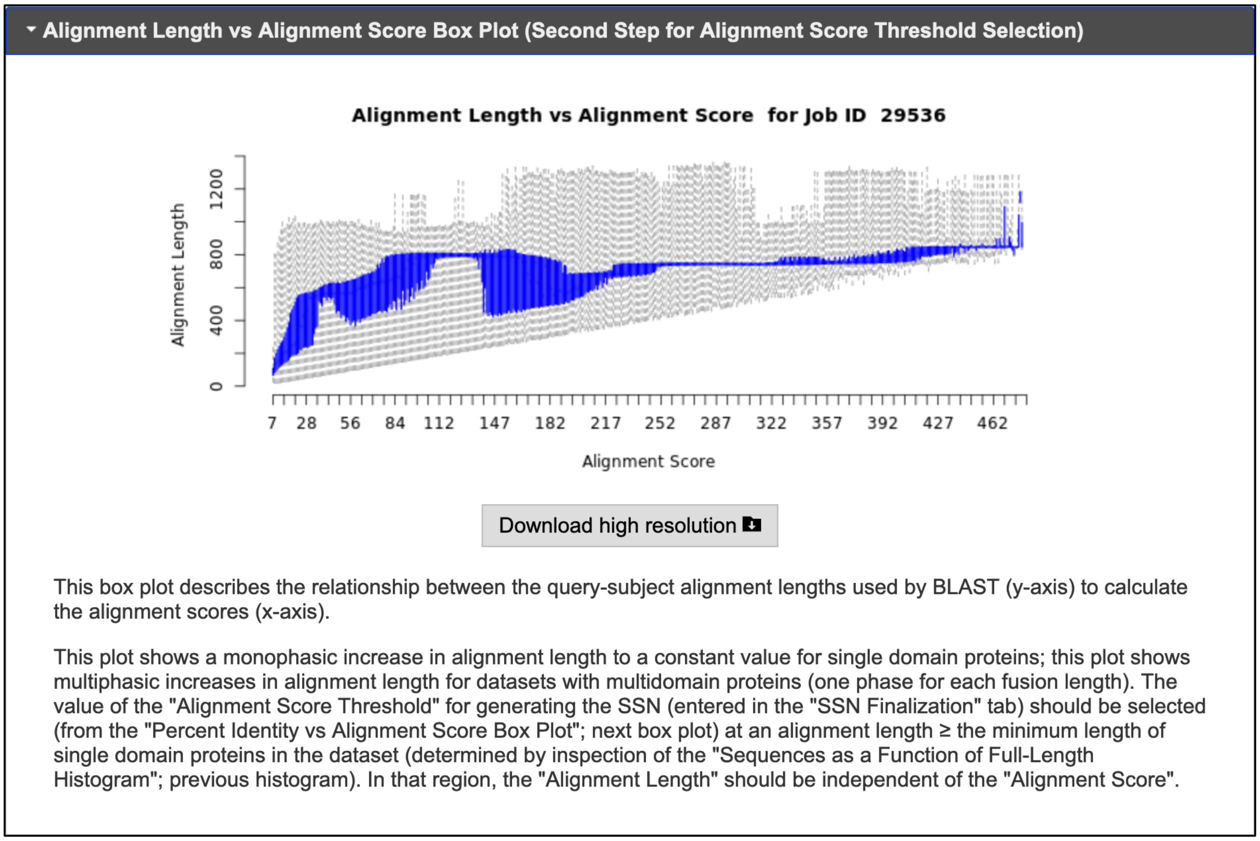
This histogram describes the length distribution for all sequences (UniProt IDs) in the input dataset.

Inspection of the histogram permits identification of fragments, single domain proteins, and multidomain fusion proteins. This histogram is used to select Minimum and Maximum "Sequence Length Restrictions" in the "SSN Finalization" tab to remove fragments, select only single domain proteins, or select multidomain proteins. The sequences in the "Sequences as a Function of Full-Length Histogram (UniRef90 Cluster IDs)" (last histogram) are used to calculate the edges.

Inspection of the histogram permits identification of fragments, single domain proteins, and multidomain fusion proteins. This histogram is used to select minimum and/or maximum sequence length restrictions to remove fragments, select single domain proteins, or select/exclude multidomain fusion proteins. These length restrictions are entered in “**Sequence Length Restrictions**” in the “**SSN Finalization**” tab of this page.

The sequences in this histogram are used to calculate the edges for the SSN; for UniRef cluster IDs, the sequences in the “**Sequences as a Function of Full-Length Histogram (UniRef Cluster IDs)**” are used to calculate the edges for the SSN.

2) “**Alignment Length vs Alignment Score Box Plot (Second Step for Alignment Score Threshold Selection)**” describes the relationship between the query-subject alignment lengths used by BLAST to calculate the alignment scores.



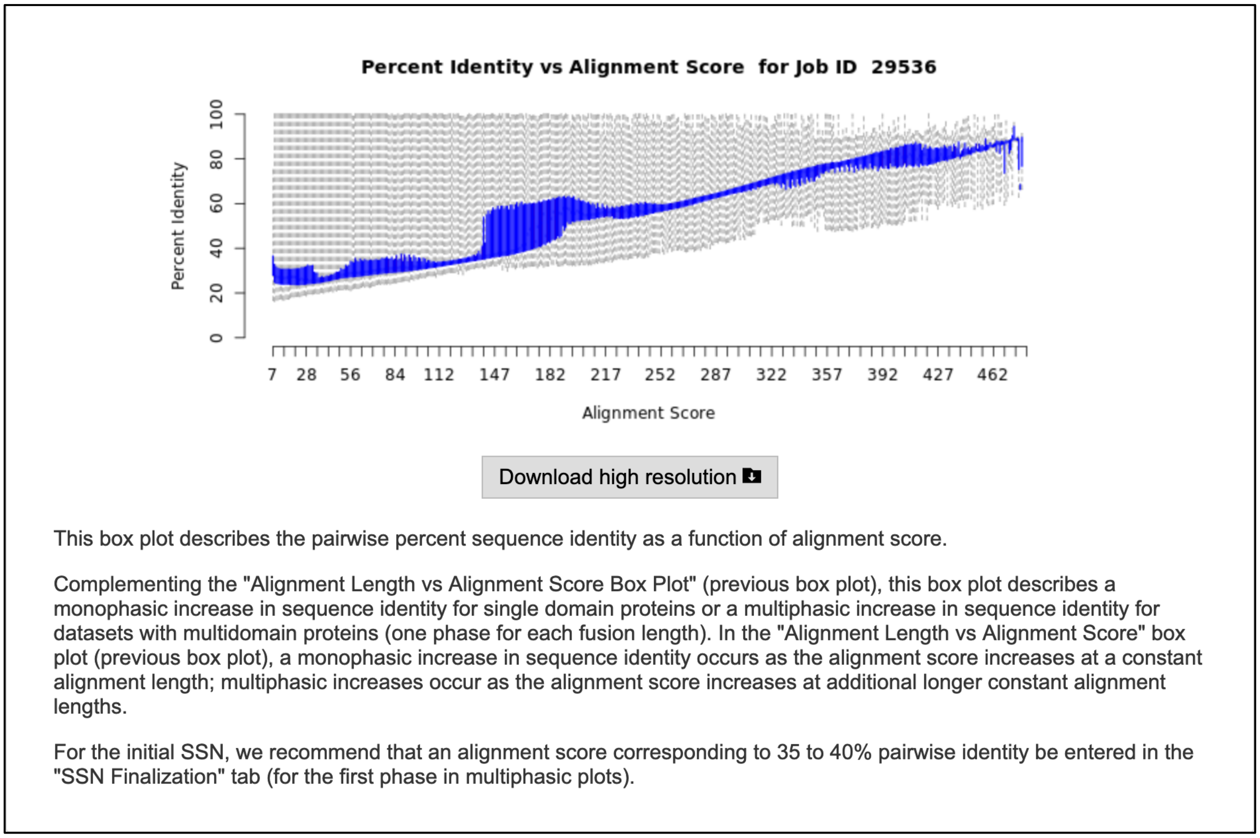
▾ Alignment Length vs Alignment Score Box Plot (Second Step for Alignment Score Threshold Selection)
Alignment Length vs Alignment Score for Job ID 29536
Alignment Length
0
400
800
1200
7
28
56
84
112
147
182
217
252
287
322
357
392
427
462
Alignment Score
Download high resolution

This box plot describes the relationship between the query-subject alignment lengths used by BLAST (y-axis) to calculate the alignment scores (x-axis).

This plot shows a monophasic increase in alignment length to a constant value for single domain proteins; this plot shows multiphasic increases in alignment length for datasets with multidomain proteins (one phase for each fusion length). The value of the "Alignment Score Threshold" for generating the SSN (entered in the "SSN Finalization" tab) should be selected (from the "Percent Identity vs Alignment Score Box Plot"; next box plot) at an alignment length ≥ the minimum length of single domain proteins in the dataset (determined by inspection of the "Sequences as a Function of Full-Length Histogram"; previous histogram). In that region, the "Alignment Length" should be independent of the "Alignment Score".

This plot shows a monophasic increase in alignment length to a constant value for single domain proteins; for datasets with multiple domain architectures, this plot shows multiphasic increases in alignment length. The minimum alignment score threshold for generating the SSN should be determined using an alignment length ≥ the minimum length for the single domain proteins (determined by inspection of the “Sequences as a Function of Length Histogram”). The actual value of the “Alignment Score Threshold” is selected using the “**Percent Identity vs Alignment Score Box Plot**”. (In the box plot, the mean values of alignment lengths are highlighted, with the “boxes” containing the 2nd and 3rd quartiles of the values. The 1st quartile of the values are shown below the box, and the 4th quartile of the values are shown above the box.)

3) “**Percent Identity vs Alignment Score Box Plot (Third Step for Alignment Score Threshold Selection)**” describes the pairwise percent sequence identity as a function of alignment score.



This box plot describes the pairwise percent sequence identity as a function of alignment score.

Complementing the "Alignment Length vs Alignment Score Box Plot" (previous box plot), this box plot describes a monophasic increase in sequence identity for single domain proteins or a multiphasic increase in sequence identity for datasets with multidomain proteins (one phase for each fusion length). In the "Alignment Length vs Alignment Score" box plot (previous box plot), a monophasic increase in sequence identity occurs as the alignment score increases at a constant alignment length; multiphasic increases occur as the alignment score increases at additional longer constant alignment lengths.

For the initial SSN, we recommend that an alignment score corresponding to 35 to 40% pairwise identity be entered in the "SSN Finalization" tab (for the first phase in multiphasic plots).

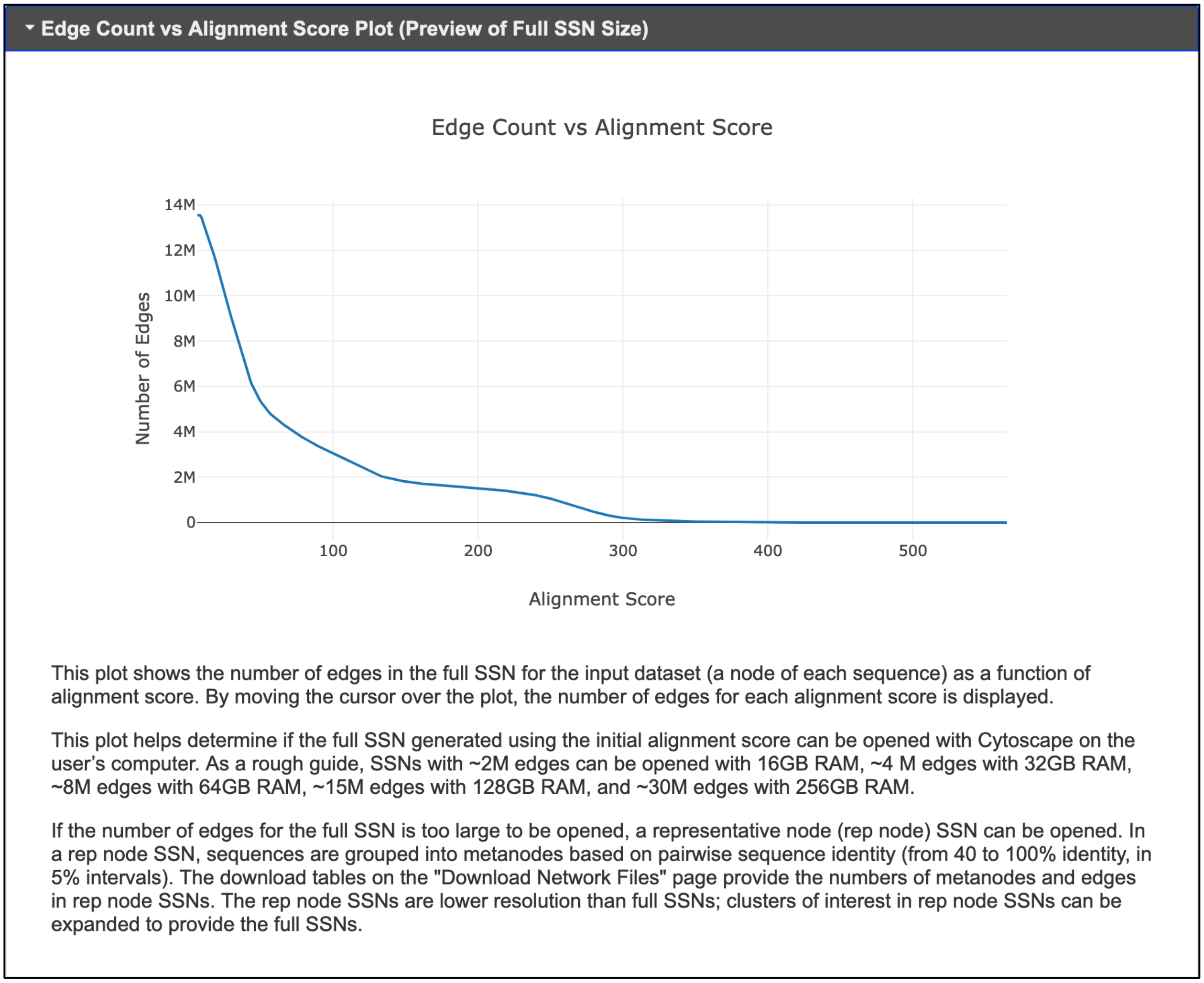
Complementing the “**Alignment Length vs Alignment Score Box Plot**”, this box plot shows a monophasic increase in sequence identity for single domain proteins or multiphasic increases in sequence identity for datasets with multiple domain fusion lengths (one phase for each fusion length). (In the box plot, the mean values of percent identity are highlighted, with the “boxes” containing the 2nd and 3rd quartiles of the values. The 1st quartile of the values are shown below the box, and the 4th quartile of the values are shown above the box.)

For the initial SSN, an alignment score corresponding to 35 to 40% pairwise identity is recommended; the selected alignment score is entered in the "**SSN Finalization**" tab (from the first phase for multiphasic datasets).

Segregation of orthologues across a functionally diverse superfamily with a single alignment score may not be possible because sequence boundaries between functions (substrate specificity and/or reaction mechanism) often do not diverge uniformly as sequence similarity decreases. Orthogonal information, e.g., genome context for microbial and fungal proteins to identify functionally linked proteins in metabolic pathways provided by EFI-GNT, can be used to assess whether clusters are isofunctional. Daughter networks for individual multifunctional clusters can be created with Cytoscape (by selecting/highlighting the clusters); larger alignment score thresholds then can be applied to a daughter network to further segregate the nodes into multiple clusters that can be assessed for isofunctionality based on SwissProt-curated/literature functions and/or genome context.

The “**Dataset Analysis**” tab also provides two additional plots for datasets generated with UniProt IDs and/or FASTA sequences:

4) “**Edge Count vs Alignment Score Plot (Preview of Full SSN Size)**” shows the number of edges in the full SSN for the input dataset (a node for each sequence) as a function of alignment score.


▾ Edge Count vs Alignment Score Plot (Preview of Full SSN Size)
Edge Count vs Alignment Score
Number of Edges
14M
12M
10M
8M
6M
4M
2M
0
100
200
300
400
500
Alignment Score

This plot shows the number of edges in the full SSN for the input dataset (a node of each sequence) as a function of alignment score. By moving the cursor over the plot, the number of edges for each alignment score is displayed.

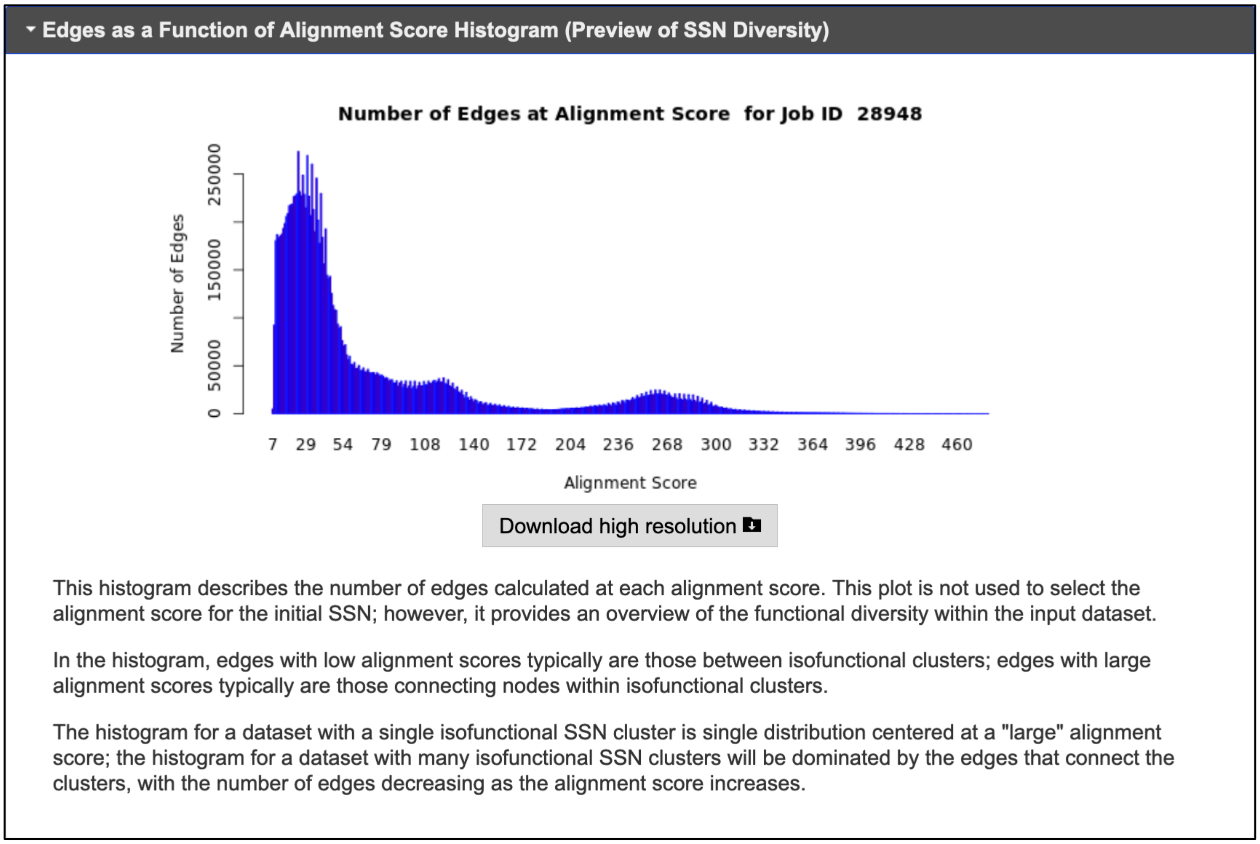
This plot helps determine if the full SSN generated using the initial alignment score can be opened with Cytoscape on the user’s computer. As a rough guide, SSNs with ~2M edges can be opened with 16GB RAM, ~4 M edges with 32GB RAM, ~8M edges with 64GB RAM, ~15M edges with 128GB RAM, and ~30M edges with 256GB RAM.

If the number of edges for the full SSN is too large to be opened, a representative node (rep node) SSN can be opened. In a rep node SSN, sequences are grouped into metanodes based on pairwise sequence identity (from 40 to 100% identity, in 5% intervals). The download tables on the "Download Network Files" page provide the numbers of metanodes and edges in rep node SSNs. The rep node SSNs are lower resolution than full SSNs; clusters of interest in rep node SSNs can be expanded to provide the full SSNs.

By moving the cursor over the plot, the number of edges for each alignment score is displayed. This plot helps determine if the full SSN generated using the initial alignment score (selected using the length histogram and sequence identity box plots) can be opened with Cytoscape on the user’s computer. The size of the SSN that can be opened is determined by the number of edges and the amount of available RAM on the user's computer; *as an approximate guide*, SSNs with ~2M edges can be opened with 16 GB RAM, ~4 M edges can be opened with 32 GB RAM, ~8M edges can be opened with 64 GB RAM, ~15M edges can be opened with 128 GB RAM, and ~30M edges can be opened with 256 GB RAM. The "**Download Network Files**" page of EFI-EST provides the number of edges in each file so the user can download the highest resolution file that can be opened.

If the number of edges for the full SSN is too large to be opened, a representative node (rep node) SSN can be opened. In a rep node SSN, sequences are grouped into metanodes based on pairwise sequence identity (from 40 to 100% identity, in 5% intervals). The download tables on the "Download Network Files" page provide the numbers of metanodes and edges in rep node SSNs. The rep node SSNs are lower resolution than full SSNs; clusters of interest in rep node SSNs can be expanded using Option D to provide complete information for selected clusters.

5) “**Edges as a Function of Alignment Score Histogram (Preview of SSN Diversity)**” provides the number of edges calculated at each alignment score.


▾ Edges as a Function of Alignment Score Histogram (Preview of SSN Diversity)
Number of Edges at Alignment Score for Job ID 28948
Number of Edges
0
50000
150000
250000
7
29
54
79
108
140
172
204
236
268
300
332
364
396
428
460
Alignment Score
Download high resolution

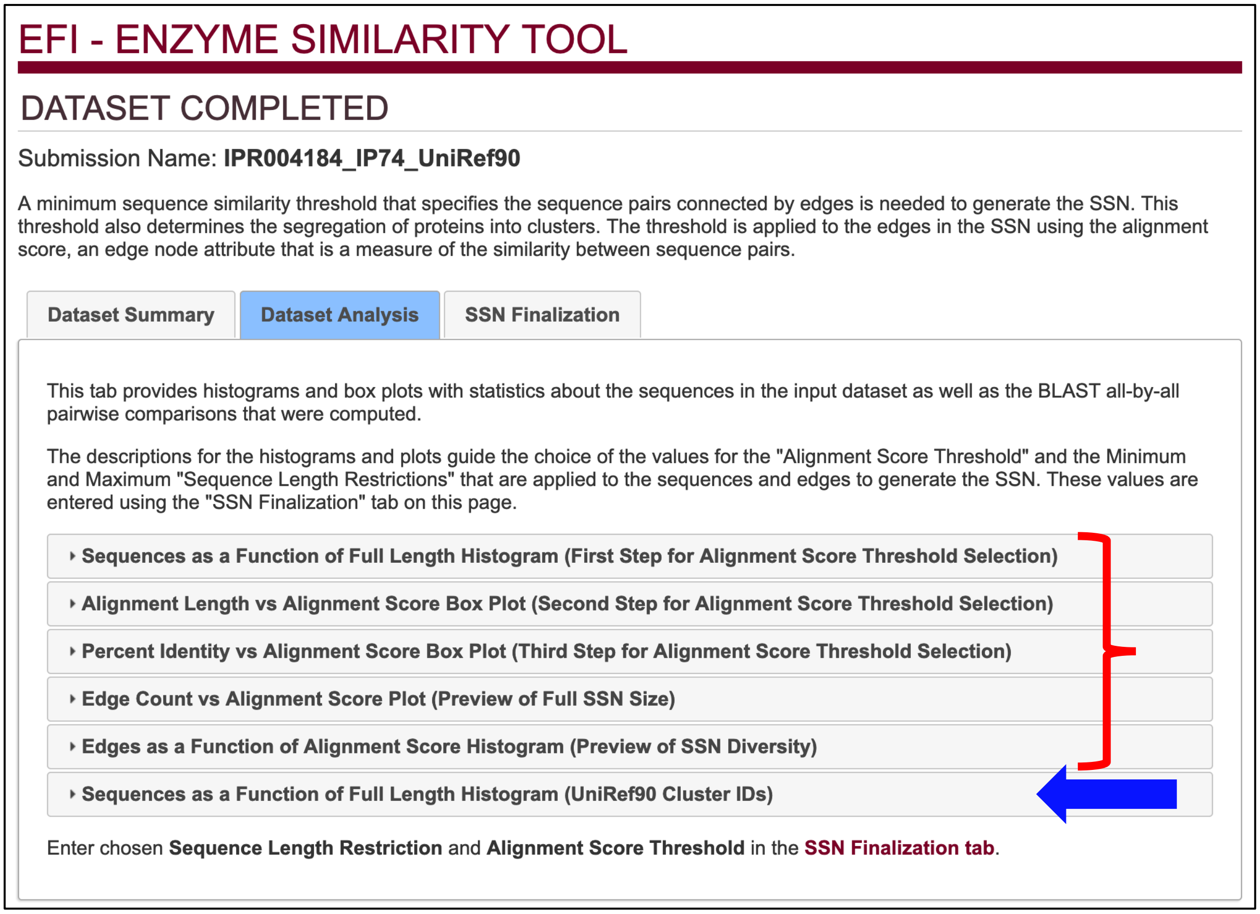
This histogram describes the number of edges calculated at each alignment score. This plot is not used to select the alignment score for the initial SSN; however, it provides an overview of the functional diversity within the input dataset.

In the histogram, edges with low alignment scores typically are those between isofunctional clusters; edges with large alignment scores typically are those connecting nodes within isofunctional clusters.

The histogram for a dataset with a single isofunctional SSN cluster is single distribution centered at a "large" alignment score; the histogram for a dataset with many isofunctional SSN clusters will be dominated by the edges that connect the clusters, with the number of edges decreasing as the alignment score increases.

This plot is not used to select the alignment score for the initial SSN; however, it provides an overview of the functional diversity within the input dataset. In the histogram, edges with low alignment scores typically are those between isofunctional clusters; edges with large alignment scores typically are those connecting nodes within isofunctional clusters. The histogram for a dataset with a single isofunctional SSN cluster is single distribution centered at a “large” alignment score; the histogram for a dataset with many isofunctional SSN clusters will be dominated by the edges that connect the clusters, with the number of edges decreasing as the alignment score increases.

Six histograms/boxplots/plots are provided when the full-length sequences of UniRef90 or UniRef50 cluster IDs are used to generate the SSN, the five provided for UniProt IDs (red bracket) and one additional for the UniRef90 or UniRef50 cluster IDs (blue arrow):



EFI - ENZYME SIMILARITY TOOL

DATASET COMPLETED

Submission Name: **IPR004184_IP74_UniRef90**

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.

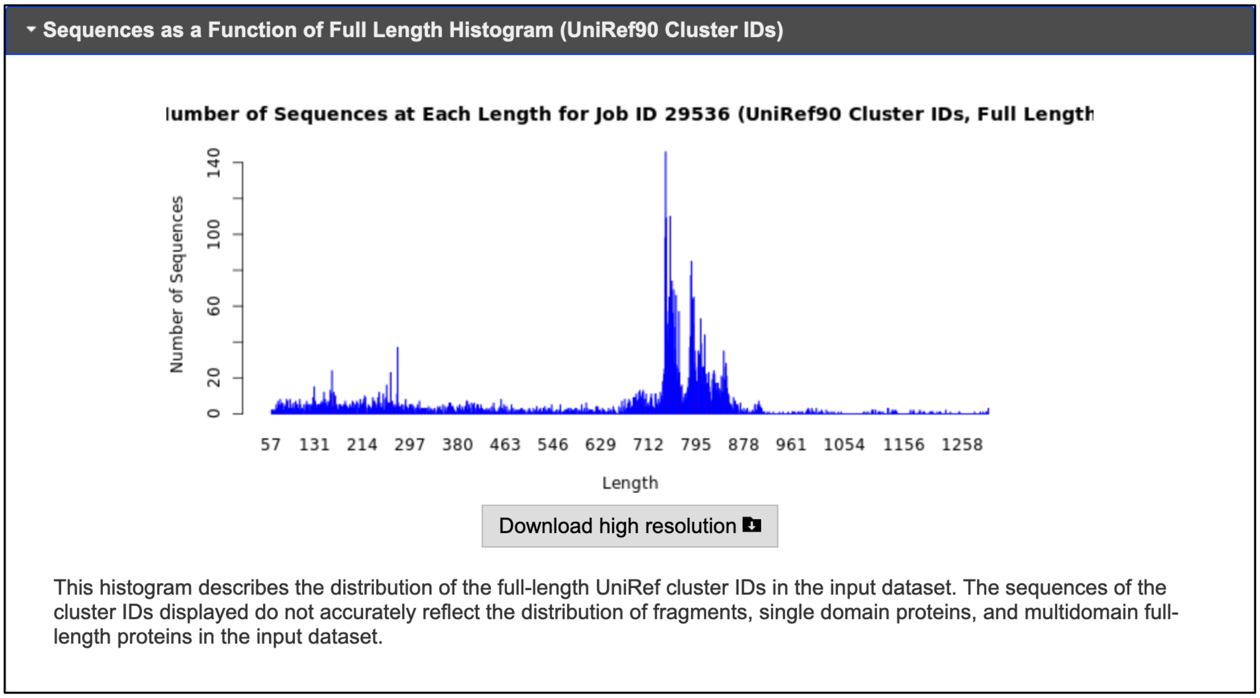
This tab provides histograms and box plots with statistics about the sequences in the input dataset as well as the BLAST all-by-all pairwise comparisons that were computed.

The descriptions for the histograms and plots guide the choice of the values for the "Alignment Score Threshold" and the Minimum and Maximum "Sequence Length Restrictions" that are applied to the sequences and edges to generate the SSN. These values are entered using the "SSN Finalization" tab on this page.

- **Sequences as a Function of Full Length Histogram (First Step for Alignment Score Threshold Selection)**
- **Alignment Length vs Alignment Score Box Plot (Second Step for Alignment Score Threshold Selection)**
- **Percent Identity vs Alignment Score Box Plot (Third Step for Alignment Score Threshold Selection)**
- **Edge Count vs Alignment Score Plot (Preview of Full SSN Size)**
- **Edges as a Function of Alignment Score Histogram (Preview of SSN Diversity)**
- **Sequences as a Function of Full Length Histogram (UniRef90 Cluster IDs)**

Enter chosen **Sequence Length Restriction** and **Alignment Score Threshold** in the **SSN Finalization tab**.

6) "**Sequences as a Function of Full-Length Histogram (UniRef Cluster IDs)**" describes the length distribution of the UniRef cluster IDs.



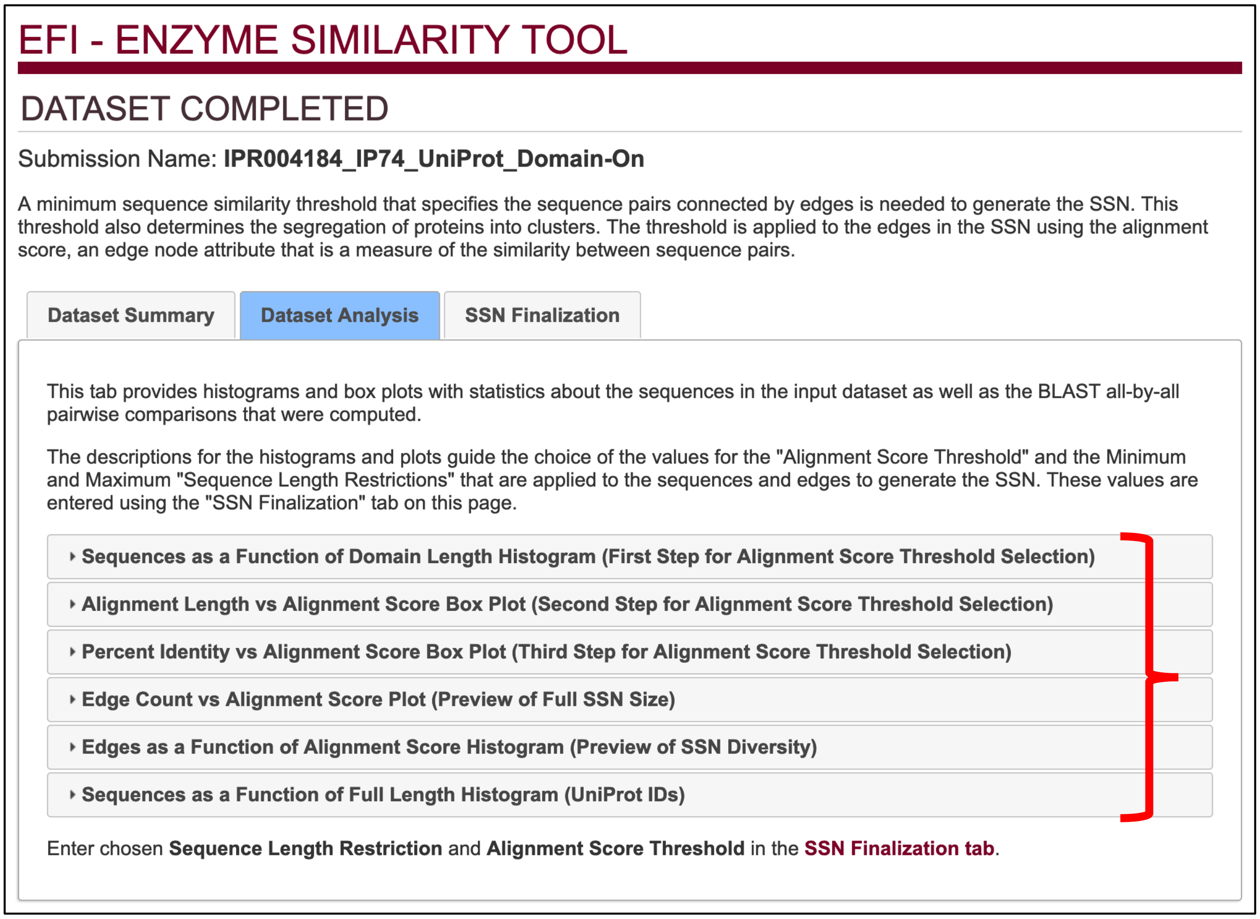
▾ Sequences as a Function of Full Length Histogram (UniRef90 Cluster IDs)

This histogram describes the distribution of the full-length UniRef cluster IDs in the input dataset. The sequences of the cluster IDs displayed do not accurately reflect the distribution of fragments, single domain proteins, and multidomain full-length proteins in the input dataset.

The sequences of the cluster IDs displayed in this histogram were used to calculate the edges for the SSN; however, the cluster IDs do not accurately reflect the distribution of fragments, single domain proteins, and multidomain proteins in the input dataset that is necessary to choose the alignment score for generating the SSN (the fragments of diverse sequence and length are over-represented). This histogram is not used to select the alignment score threshold but is provided for information purposes; the length distribution of the total set of sequences in "**Sequences as a Function of Full-Length Histogram (First Step for Alignment Score Threshold Selection)**" should be used to choose the minimum alignment score threshold.

2. “Dataset Analysis” tab for domain sequences (Options B and D)

Six histograms/boxplots/plots (red bracket) are provided when domains in UniProt IDs are used to generate the SSN.



EFI - ENZYME SIMILARITY TOOL

DATASET COMPLETED

Submission Name: **IPR004184_IP74_UniProt_Domain-On**

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.

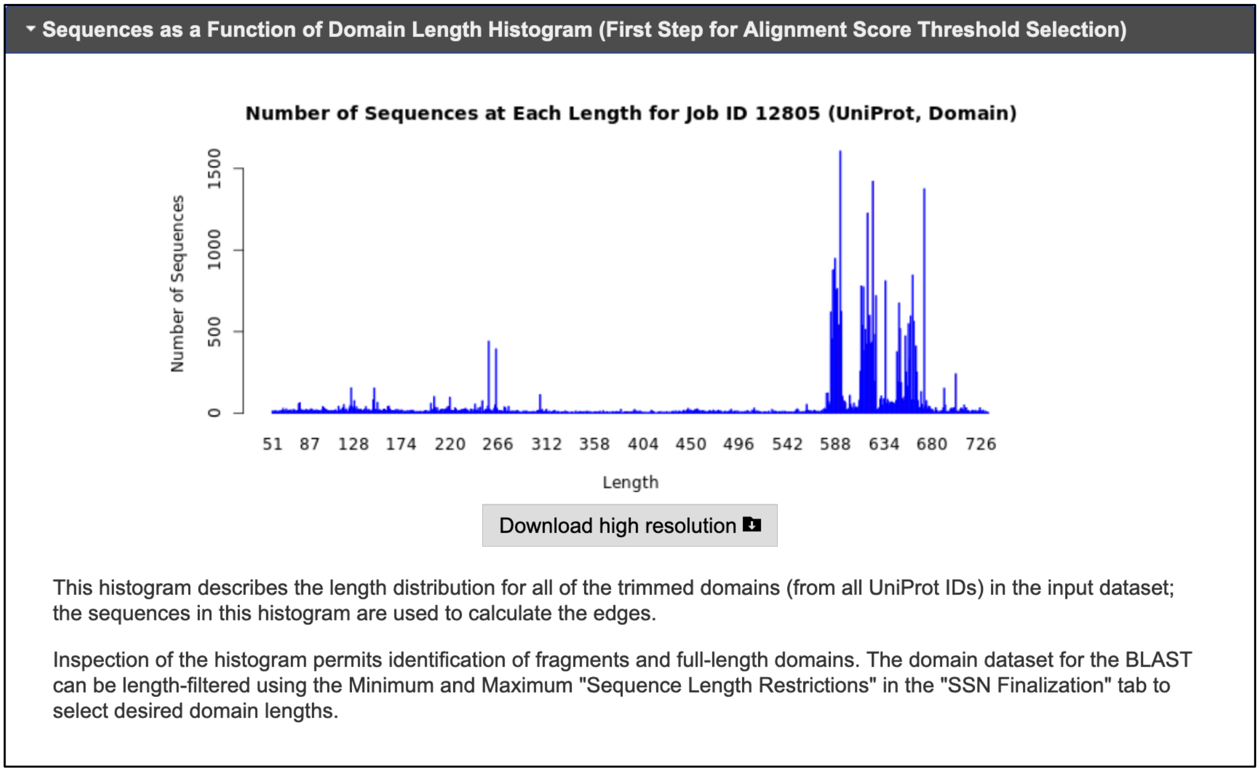
This tab provides histograms and box plots with statistics about the sequences in the input dataset as well as the BLAST all-by-all pairwise comparisons that were computed.

The descriptions for the histograms and plots guide the choice of the values for the "Alignment Score Threshold" and the Minimum and Maximum "Sequence Length Restrictions" that are applied to the sequences and edges to generate the SSN. These values are entered using the "SSN Finalization" tab on this page.

- **Sequences as a Function of Domain Length Histogram (First Step for Alignment Score Threshold Selection)**
- **Alignment Length vs Alignment Score Box Plot (Second Step for Alignment Score Threshold Selection)**
- **Percent Identity vs Alignment Score Box Plot (Third Step for Alignment Score Threshold Selection)**
- **Edge Count vs Alignment Score Plot (Preview of Full SSN Size)**
- **Edges as a Function of Alignment Score Histogram (Preview of SSN Diversity)**
- **Sequences as a Function of Full Length Histogram (UniProt IDs)**

Enter chosen **Sequence Length Restriction** and **Alignment Score Threshold** in the **SSN Finalization tab**.

1) “**Sequences as a Function of Domain-Length Histogram (First Step for Alignment Score Threshold Selection)**” describes the length distribution of all trimmed domain sequences, e.g., from all UniProt IDs, in the input dataset.



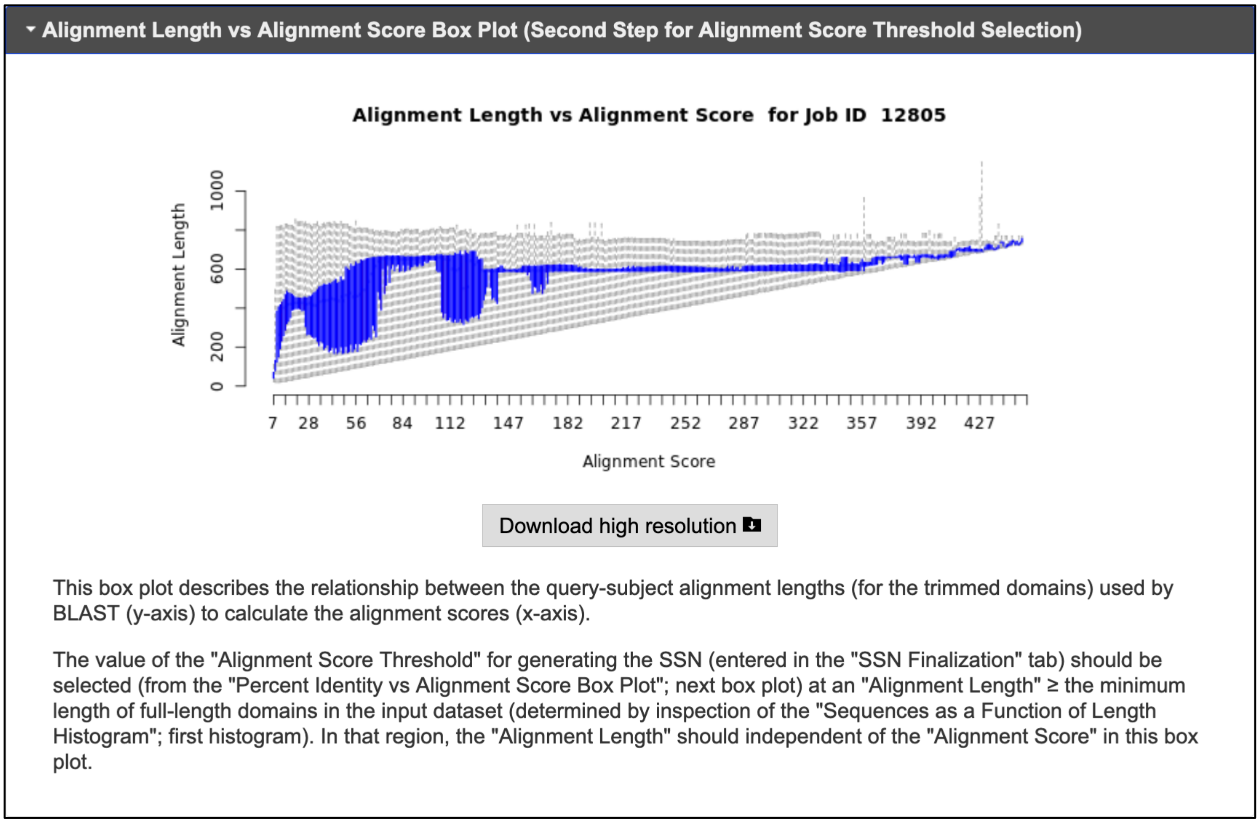
▾ Sequences as a Function of Domain Length Histogram (First Step for Alignment Score Threshold Selection)

This histogram describes the length distribution for all of the trimmed domains (from all UniProt IDs) in the input dataset; the sequences in this histogram are used to calculate the edges.

Inspection of the histogram permits identification of fragments and full-length domains. The domain dataset for the BLAST can be length-filtered using the Minimum and Maximum "Sequence Length Restrictions" in the "SSN Finalization" tab to select desired domain lengths.

Inspection of the histogram permits identification of fragments, e.g., if the domains defined by the single Pfam family or InterPro family/domain used to generate the SSN are interrupted by insertion sequences, the partial domains will appear in this histogram as fragments. The domain dataset can be length-filtered using the “**Sequence Length Restrictions**” in the “**SSN Finalization**” tab to remove the fragments from the SSN.

2) “**Alignment Length vs Alignment Score Box Plot (Second Step for Alignment Score Threshold Selection)**” describes the relationship between the query-subject alignment lengths used by BLAST to calculate the alignment scores.



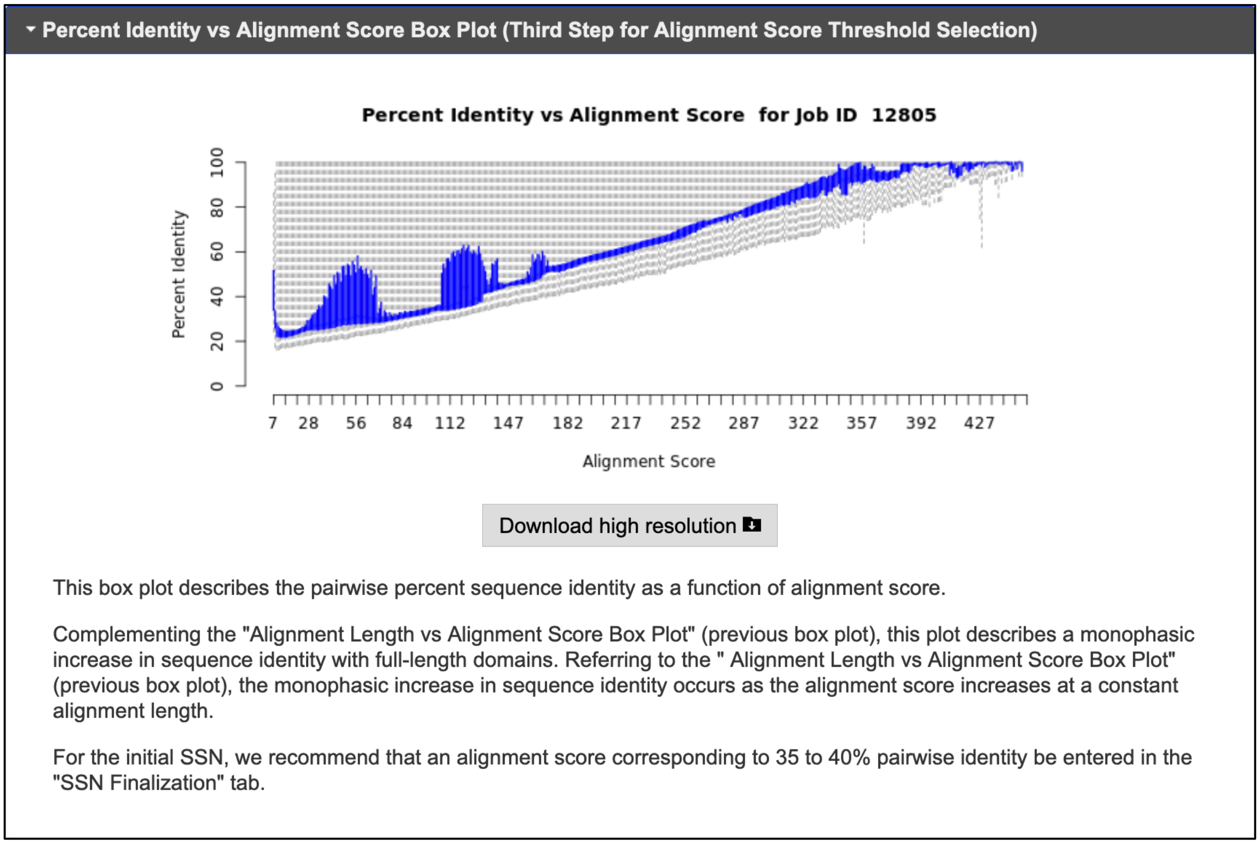
▾ Alignment Length vs Alignment Score Box Plot (Second Step for Alignment Score Threshold Selection)

This box plot describes the relationship between the query-subject alignment lengths (for the trimmed domains) used by BLAST (y-axis) to calculate the alignment scores (x-axis).

The value of the "Alignment Score Threshold" for generating the SSN (entered in the "SSN Finalization" tab) should be selected (from the "Percent Identity vs Alignment Score Box Plot"; next box plot) at an "Alignment Length" ≥ the minimum length of full-length domains in the input dataset (determined by inspection of the "Sequences as a Function of Length Histogram"; first histogram). In that region, the "Alignment Length" should independent of the "Alignment Score" in this box plot.

This plot should show a monophasic increase in alignment length to a constant value for the domain length. The minimum alignment score threshold for generating the SSN should be determined using an alignment length that is ≥ the minimum length for the domain [determined by inspection of the “**Sequences as a Function of Domain Length Histogram (First Step for Alignment Score Threshold Selection)**”; previous histogram]. The actual value of the “**Alignment Score Threshold**” entered in the “**SSN Finalization**” tab is selected using the “**Percent Identity vs Alignment Score Box Plot (Third Step for Alignment Score Threshold Selection)**” (next box plot). (In the box plot, the mean values of alignment lengths are highlighted, with the “boxes” containing the 2nd and 3rd quartiles of the values. The 1st quartile of the values are shown below the box, and the 4th quartile of the values are shown above the box.)

3) “**Percent Identity vs Alignment Score Box Plot (Third Step for Alignment Score Threshold Selection)**” describes the pairwise percent sequence identity as a function of alignment score.



▾ Percent Identity vs Alignment Score Box Plot (Third Step for Alignment Score Threshold Selection)

This box plot describes the pairwise percent sequence identity as a function of alignment score.

Complementing the "Alignment Length vs Alignment Score Box Plot" (previous box plot), this plot describes a monophasic increase in sequence identity with full-length domains. Referring to the " Alignment Length vs Alignment Score Box Plot" (previous box plot), the monophasic increase in sequence identity occurs as the alignment score increases at a constant alignment length.

For the initial SSN, we recommend that an alignment score corresponding to 35 to 40% pairwise identity be entered in the "SSN Finalization" tab.

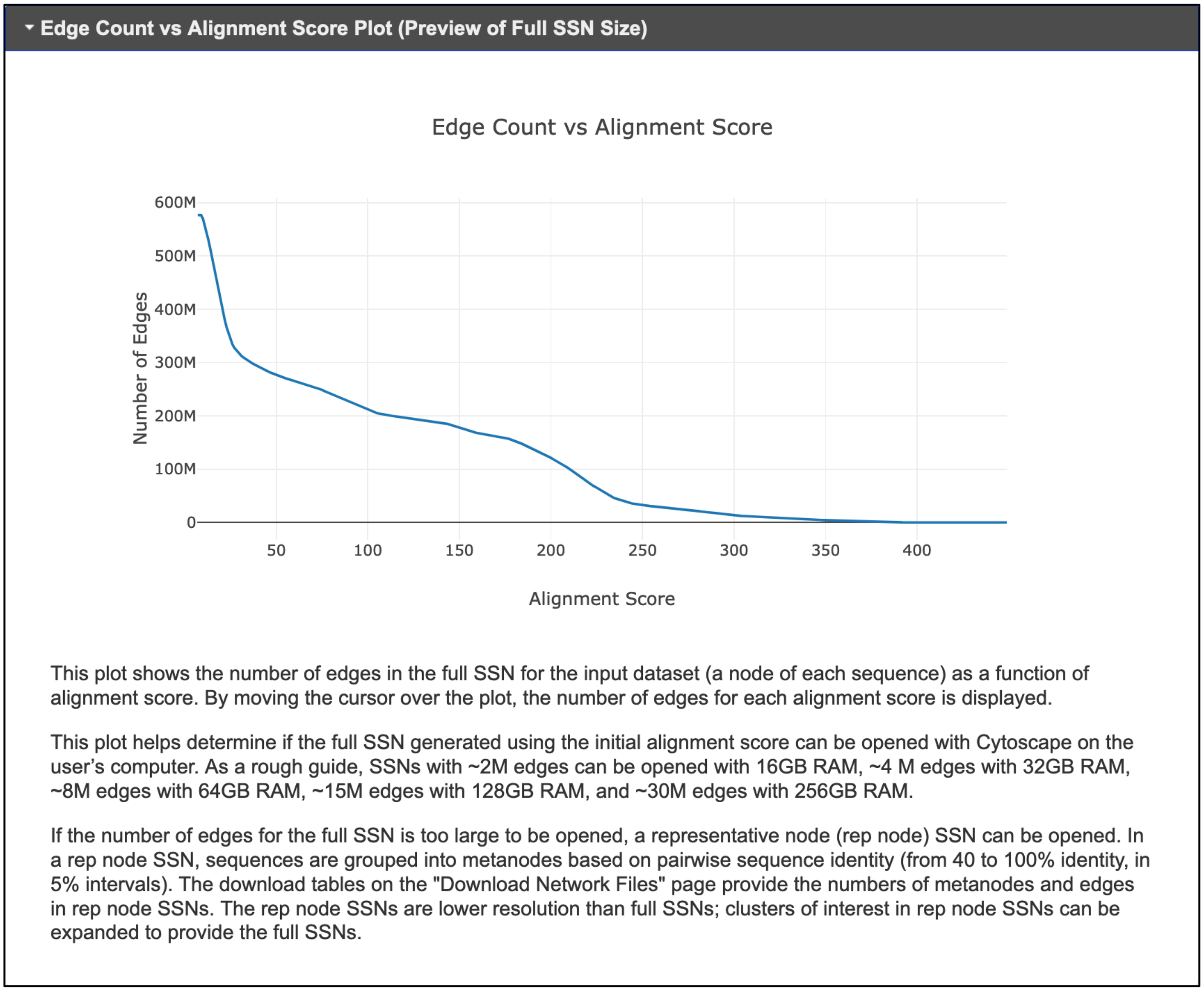
Complementing the “**Alignment Length vs Alignment Score Box Plot (Third Step for Alignment Score Threshold Selection)**”, this box plot should show a monophasic increase in sequence identity. (In the box plot, the mean values of percent identity are highlighted, with the “boxes” containing the 2nd and 3rd quartiles of the values. The 1st quartile of the values are shown below the box, and the 4th quartile of the values are shown above the box.)

For the initial SSN, an alignment score corresponding to 35 to 40% pairwise identity is recommended; the selected alignment score is entered in the "**SSN Finalization**" tab (from the first phase for multiphasic datasets).

Segregation of orthologues across a functionally diverse superfamily with a single alignment score may not be possible because sequence boundaries between functions (substrate specificity and/or reaction mechanism) often do not diverge uniformly as sequence similarity decreases. Orthogonal information, e.g., genome context for microbial and fungal proteins to identify functionally linked proteins in metabolic pathways provided by EFI-GNT, can be used to assess whether clusters are isofunctional. Daughter networks for individual multifunctional clusters can be created with Cytoscape (by selecting/highlighting the clusters); larger alignment score thresholds then can be applied to a daughter network to further segregate the nodes into multiple clusters that can be assessed for isofunctionality based on SwissProt-curated/literature functions and/or genome context.

The "**Dataset Analysis**" tab also provides three additional plots for datasets generated with UniProt IDs:

4) "**Edge Count vs Alignment Score Plot (Preview of Full SSN Size)**" shows the number of edges in the full SSN for the input dataset (a node of each sequence) as a function of alignment score.


▾ Edge Count vs Alignment Score Plot (Preview of Full SSN Size)
Edge Count vs Alignment Score
Number of Edges
600M
500M
400M
300M
200M
100M
0
50
100
150
200
250
300
350
400
Alignment Score

This plot shows the number of edges in the full SSN for the input dataset (a node of each sequence) as a function of alignment score. By moving the cursor over the plot, the number of edges for each alignment score is displayed.

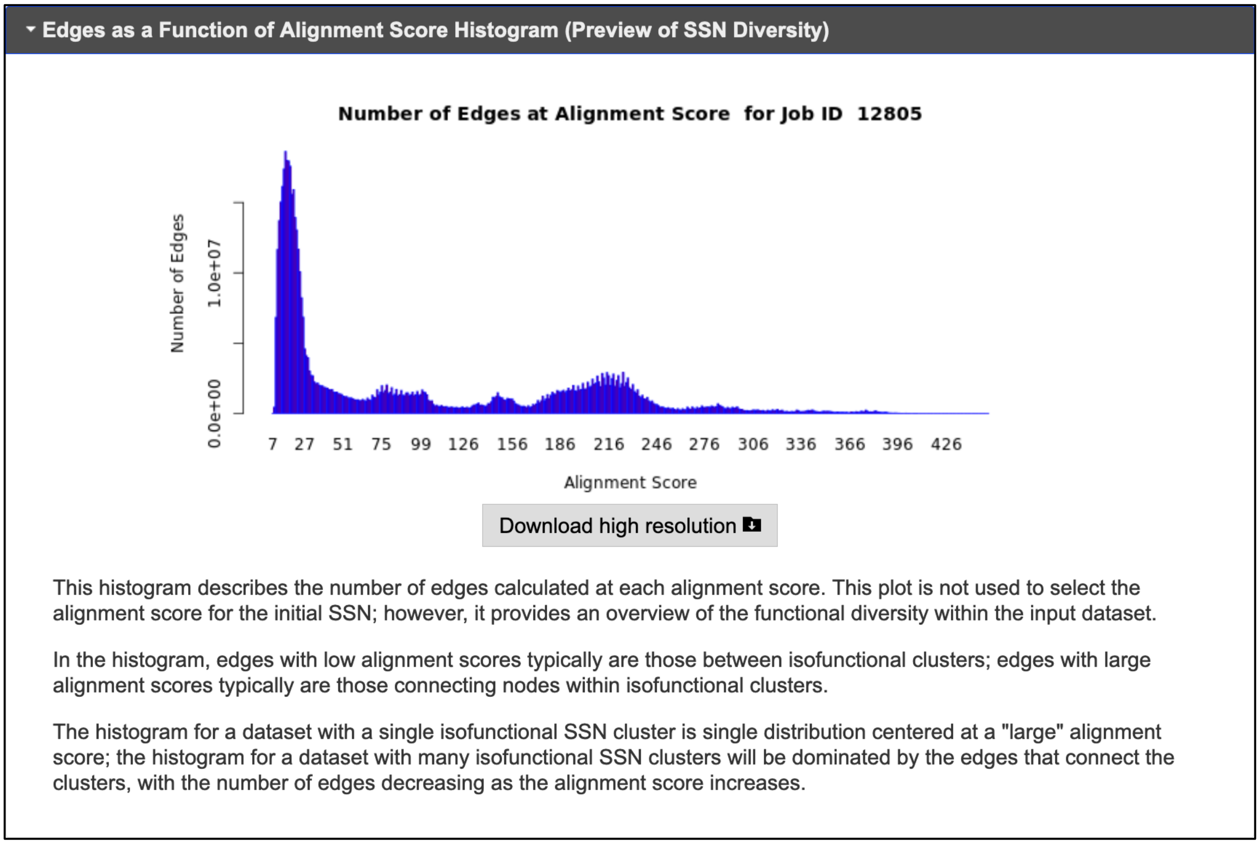
This plot helps determine if the full SSN generated using the initial alignment score can be opened with Cytoscape on the user's computer. As a rough guide, SSNs with ~2M edges can be opened with 16GB RAM, ~4 M edges with 32GB RAM, ~8M edges with 64GB RAM, ~15M edges with 128GB RAM, and ~30M edges with 256GB RAM.

If the number of edges for the full SSN is too large to be opened, a representative node (rep node) SSN can be opened. In a rep node SSN, sequences are grouped into metanodes based on pairwise sequence identity (from 40 to 100% identity, in 5% intervals). The download tables on the "Download Network Files" page provide the numbers of metanodes and edges in rep node SSNs. The rep node SSNs are lower resolution than full SSNs; clusters of interest in rep node SSNs can be expanded to provide the full SSNs.

By moving the cursor over the plot, the number of edges for each alignment score is displayed. This plot helps determine if the full SSN generated using the initial alignment score (selected using the length histogram and sequence identity box plots) can be opened with Cytoscape on the user's computer. The size of the SSN that can be opened is determined by the number of edges and the amount of available RAM on the user's computer; *as an approximate guide*, SSNs with ~2M edges can be opened with 16 GB RAM, ~4 M edges can be opened with 32 GB RAM, ~8M edges can be opened with 64 GB RAM, ~15M edges can be opened with 128 GB RAM, and ~30M edges can be opened with 256 GB RAM. The "**Download Network Files**" page of EFI-EST provides the number of edges in each file so the user can download the highest resolution file that can be opened.

If the number of edges for the full SSN is too large to be opened, a representative node (rep node) SSN can be opened. In a rep node SSN, sequences are grouped into metanodes based on pairwise sequence identity (from 40 to 100% identity, in 5% intervals). The download tables on the "Download Network Files" page provide the numbers of metanodes and edges in rep node SSNs. The rep node SSNs are lower resolution than full SSNs; clusters of interest in rep node SSNs can be expanded to provide the full SSNs.

5) “**Edges as a Function of Alignment Score Histogram (Preview of SSN Diversity)**” provides the number of edges calculated at each alignment score. This plot is not used to select the alignment score for the initial SSN; however, it provides an overview of the functional diversity within the input dataset.



▾ Edges as a Function of Alignment Score Histogram (Preview of SSN Diversity)

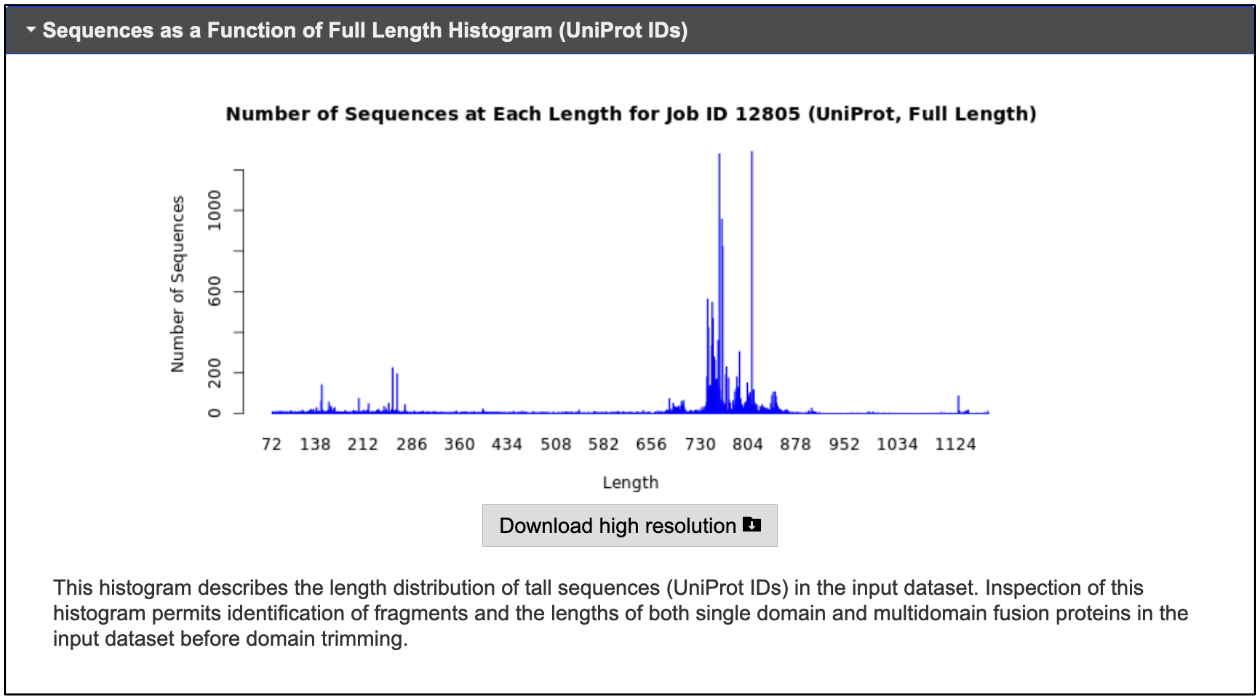
This histogram describes the number of edges calculated at each alignment score. This plot is not used to select the alignment score for the initial SSN; however, it provides an overview of the functional diversity within the input dataset.

In the histogram, edges with low alignment scores typically are those between isofunctional clusters; edges with large alignment scores typically are those connecting nodes within isofunctional clusters.

The histogram for a dataset with a single isofunctional SSN cluster is single distribution centered at a "large" alignment score; the histogram for a dataset with many isofunctional SSN clusters will be dominated by the edges that connect the clusters, with the number of edges decreasing as the alignment score increases.

In the histogram, edges with low alignment scores typically are those between isofunctional clusters; edges with large alignment scores typically are those connecting nodes within isofunctional clusters. The histogram for a dataset with a single isofunctional SSN cluster is single distribution centered at a “large” alignment score; the histogram for a dataset with many isofunctional SSN clusters will be dominated by the edges that connect the clusters, with the number of edges decreasing as the alignment score increases.

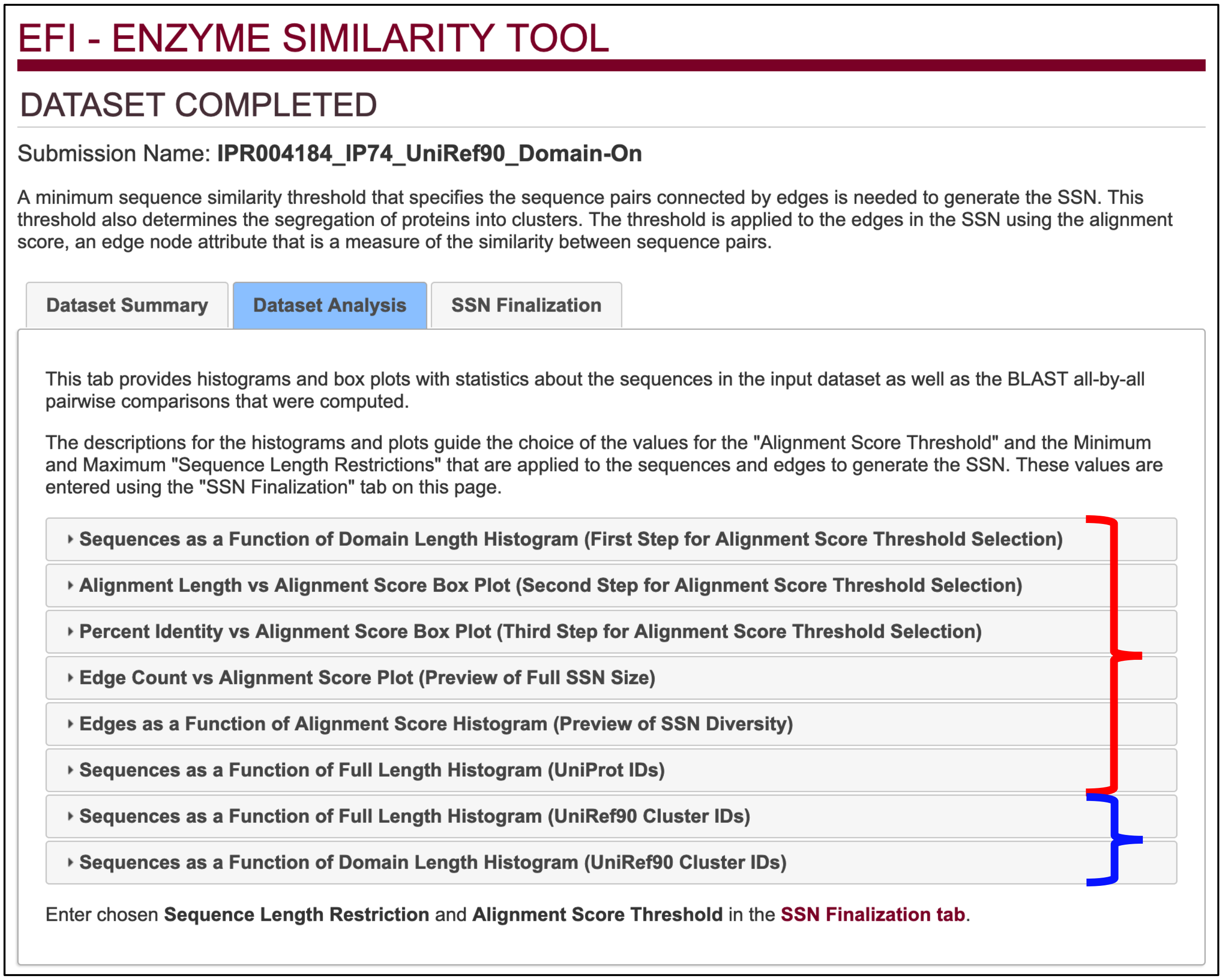
6) “**Sequences as a Function of Full-Length Histogram (UniProt IDs)**” describes the full-length distribution for all UniProt IDs in the input dataset before domain trimming. This histogram is not used to select the alignment score threshold but is provided for information purposes.



▾ Sequences as a Function of Full Length Histogram (UniProt IDs)

This histogram describes the length distribution of tall sequences (UniProt IDs) in the input dataset. Inspection of this histogram permits identification of fragments and the lengths of both single domain and multidomain fusion proteins in the input dataset before domain trimming.

Eight histograms/boxplots/plots are provided when UniRef90 or UniRef50 cluster IDs are used to generate the SSN, the six provided for UniProt (red bracket) and two additional for the UniRef90 or UniRef50 cluster IDs (blue bracket):



EFI - ENZYME SIMILARITY TOOL

DATASET COMPLETED

Submission Name: **IPR004184_IP74_UniRef90_Domain-On**

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.

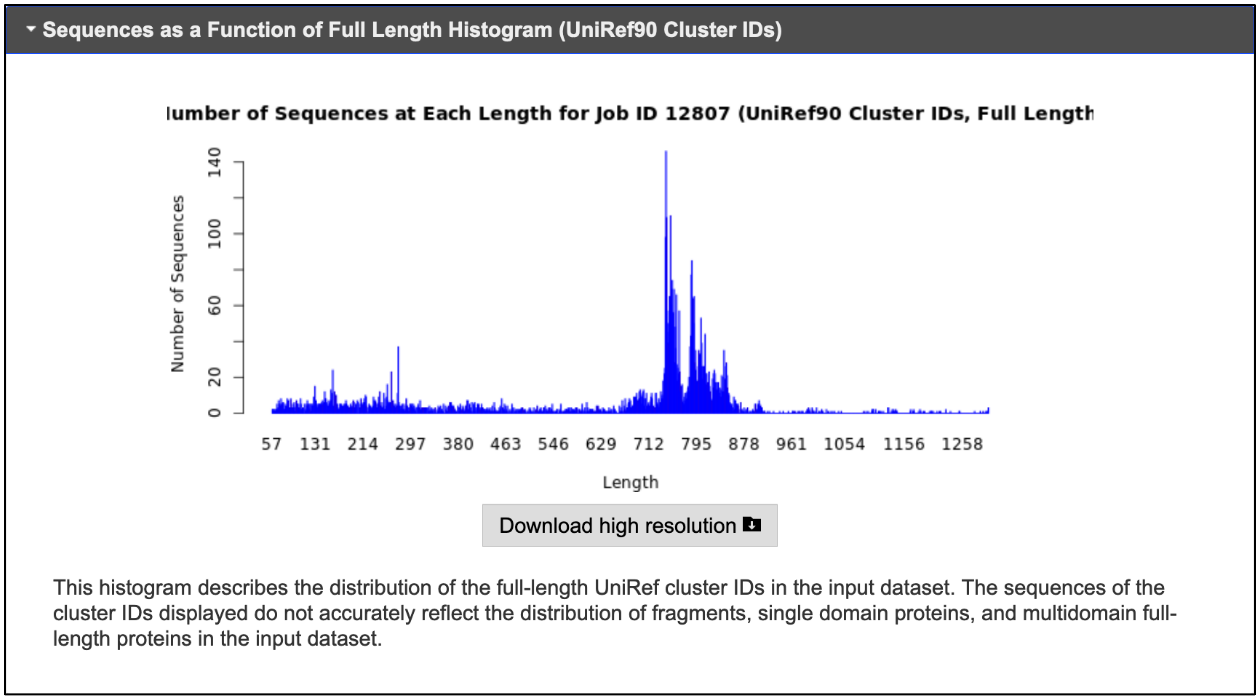
This tab provides histograms and box plots with statistics about the sequences in the input dataset as well as the BLAST all-by-all pairwise comparisons that were computed.

The descriptions for the histograms and plots guide the choice of the values for the "Alignment Score Threshold" and the Minimum and Maximum "Sequence Length Restrictions" that are applied to the sequences and edges to generate the SSN. These values are entered using the "SSN Finalization" tab on this page.

- **Sequences as a Function of Domain Length Histogram (First Step for Alignment Score Threshold Selection)**
- **Alignment Length vs Alignment Score Box Plot (Second Step for Alignment Score Threshold Selection)**
- **Percent Identity vs Alignment Score Box Plot (Third Step for Alignment Score Threshold Selection)**
- **Edge Count vs Alignment Score Plot (Preview of Full SSN Size)**
- **Edges as a Function of Alignment Score Histogram (Preview of SSN Diversity)**
- **Sequences as a Function of Full Length Histogram (UniProt IDs)**
- **Sequences as a Function of Full Length Histogram (UniRef90 Cluster IDs)**
- **Sequences as a Function of Domain Length Histogram (UniRef90 Cluster IDs)**

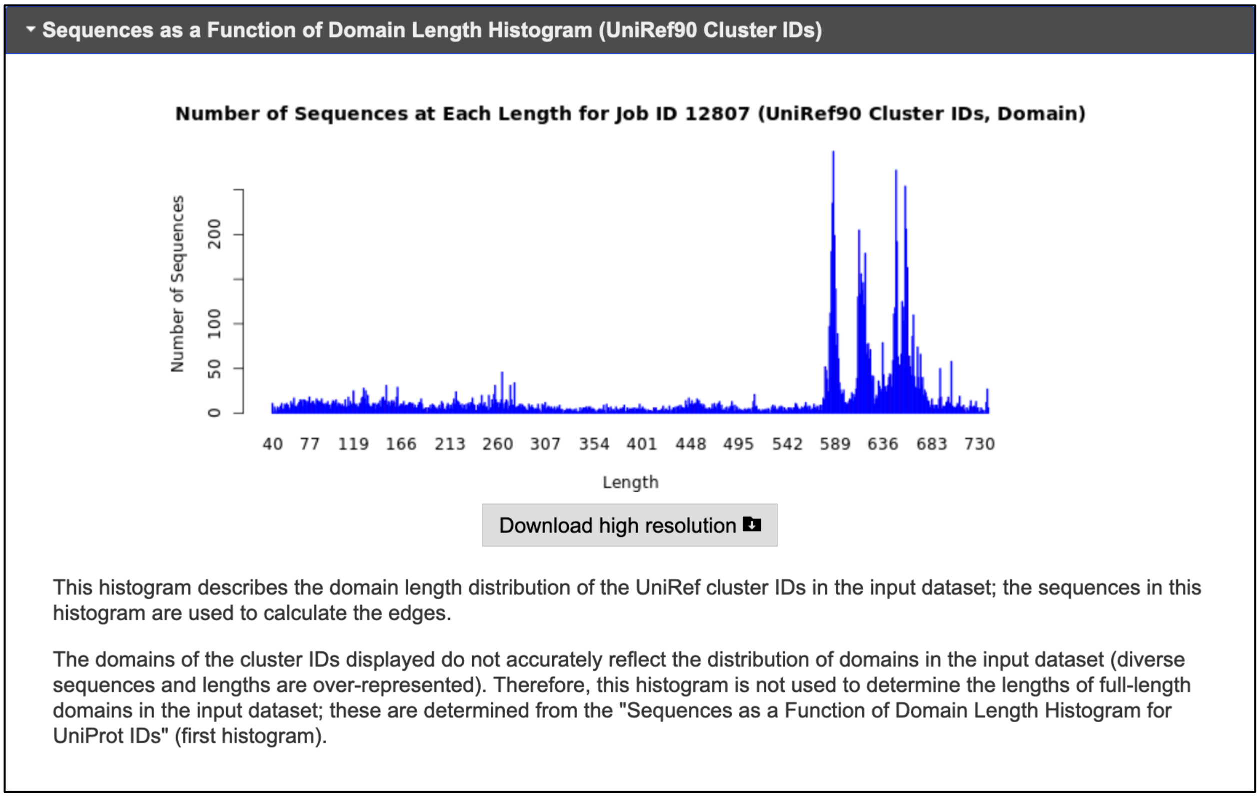
Enter chosen **Sequence Length Restriction** and **Alignment Score Threshold** in the **SSN Finalization tab**.

7) “**Sequences as a Function of Full-Length Histogram (UniRef Cluster IDs)**” describes the full-length distribution of the UniRef cluster IDs. This histogram is not used to select the alignment score threshold but is provided for information purposes.


▾ Sequences as a Function of Full Length Histogram (UniRef90 Cluster IDs)
lumber of Sequences at Each Length for Job ID 12807 (UniRef90 Cluster IDs, Full Length
Number of Sequences
0
20
60
100
140
57
131
214
297
380
463
546
629
712
795
878
961
1054
1156
1258
Length
Download high resolution

This histogram describes the distribution of the full-length UniRef cluster IDs in the input dataset. The sequences of the cluster IDs displayed do not accurately reflect the distribution of fragments, single domain proteins, and multidomain full-length proteins in the input dataset.

8) “**Sequences as a Function of Domain-Length Histogram (Domain Length for UniRef Cluster IDs)**” describes the length distribution of the UniRef cluster ID domains.



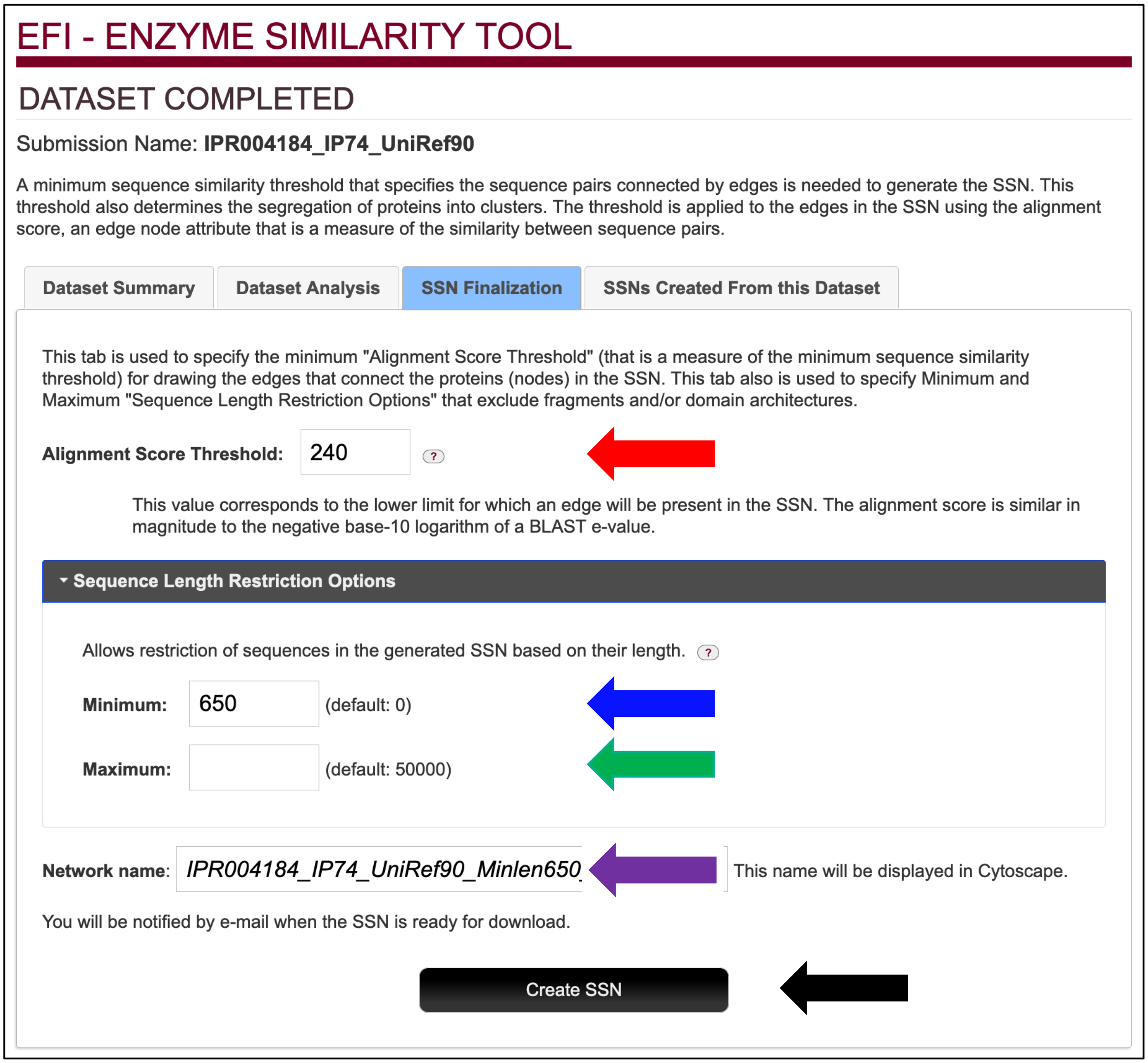
▾ Sequences as a Function of Domain Length Histogram (UniRef90 Cluster IDs)

This histogram describes the domain length distribution of the UniRef cluster IDs in the input dataset; the sequences in this histogram are used to calculate the edges.

The domains of the cluster IDs displayed do not accurately reflect the distribution of domains in the input dataset (diverse sequences and lengths are over-represented). Therefore, this histogram is not used to determine the lengths of full-length domains in the input dataset; these are determined from the "Sequences as a Function of Domain Length Histogram for UniProt IDs" (first histogram).

The sequences of the cluster ID domains displayed in this histogram were used to calculate the edges for the SSN, although they do not accurately reflect the distribution of domains in the input dataset necessary to choose the alignment score for generating the SSN (fragments of diverse sequence and length are over-represented). This histogram is not used to select the alignment score threshold but is provided for information purposes; the length distribution of the total set of domain sequences in the “**Sequences as a Function of Length Histogram (First Step for Alignment Score Threshold Selection)**” should be used to choose the minimum alignment score threshold.

"SSN Finalization" tab. This tab is used to input the minimum "**Alignment Score Threshold**" and minimum and maximum "**Sequence Length Restriction Options**" for generating the SSN determined using the histograms and box plots provided in the "**Dataset Analysis**" tab.



EFI - ENZYME SIMILARITY TOOL

DATASET COMPLETED

Submission Name: **IPR004184_IP74_UniRef90**

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. This tab also is used to specify Minimum and Maximum "Sequence Length Restriction Options" that exclude fragments and/or domain architectures.

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: *IPR004184_IP74_UniRef90_Minlen650* This name will be displayed in Cytoscape.

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

The alignment score selected as described in the description of the "**Dataset Analysis**" tab is entered in the "**Alignment Score Threshold**" box red arrow; required). In the "**Sequence Length Restrictions Options**", the user can enter (optional) "**Minimum**" (blue arrow) and/or

"**Maximum**" (green arrow) length filters to remove fragments, include single domain proteins, or exclude multidomain proteins. For SSNs generated with UniProt IDs, the filters are applied to all sequences. For SSNs generated with UniRef90 or UniRef50 cluster IDs or domains, the filters are applied to both the cluster IDs as well as the sequences in the clusters. First, the sequences in each UniRef cluster are length-filtered; those that do not satisfy the requirements are removed from the cluster. Then, the UniRef cluster ID sequences are length-filtered; those that do not satisfy the criteria are discarded. A UniRef cluster may contain sequences longer than the cluster ID (e.g., the seed sequence used to identify the cluster) that exceed the minimum length filter even if the cluster ID sequence does not; however, because UniRef clusters are discarded based the length of their cluster IDs, these "acceptable" sequences are discarded.

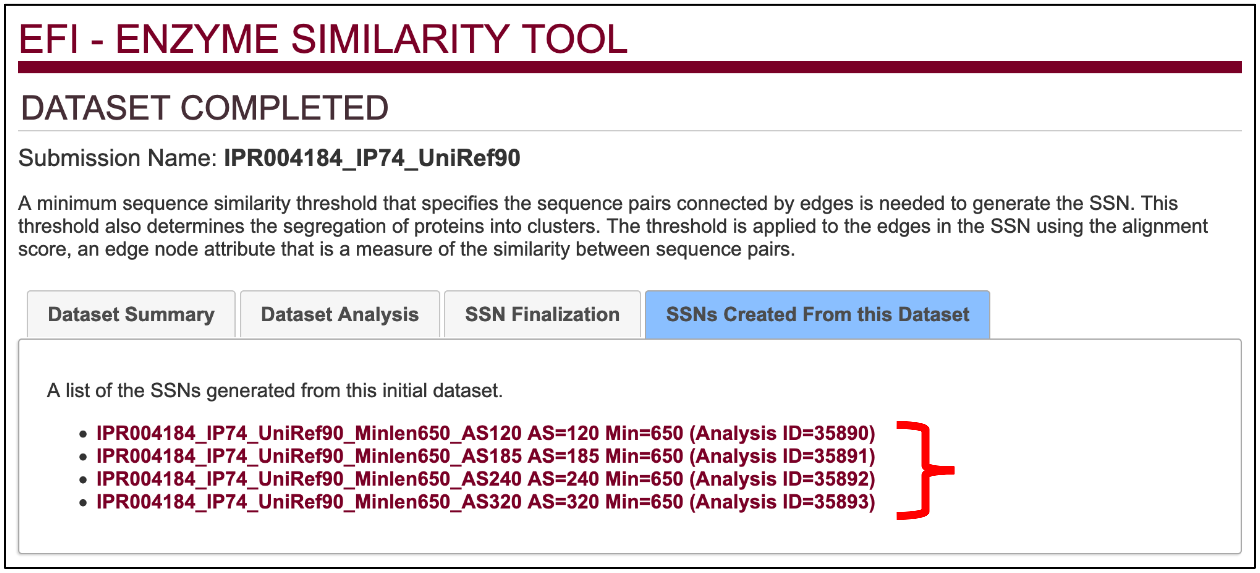
Typically, the minimum length filter is used to eliminate fragments; when desired, the maximum length filter is used to exclude multidomain proteins. Alternatively, the minimum and maximum length filters can be used together to select multidomain proteins (for full-length proteins for Options A, B, C, and D). The presence of fragments contained within UniRef clusters will not affect the SSNs (generated with the cluster IDs) nor the GNNs (although the statistics will be based on neighbors only on one side of the query gene because fragments often occur because the query gene is at the end of a sequencing contig). However, fragments will influence the consensus sequences for ShortBRED families in CGFP and, therefore, the sequence motifs that are used to determine metagenome abundance. The user can choose to not apply length filters for generating the SSN. If the user subsequently decided to submit the SSN to EFI-CGFP, the "**Run CGFP/ShortBRED**" tab on the EFI-CGFP home page provides the same "**Sequence Length**

Restriction Options" so that fragments included in UniRef clusters in the SSNs can be removed for CGFP marker identification and metagenome abundance quantification.

A "**Network name**" panel (magenta arrow; will be prefilled with the job name entered in the Option A, B, C, and D home pages) is required that will appear in the Cytoscape session and be used on the "**Previous Jobs**" panel on the EFI-EST home page and the "**Job History**" page; this name also will appear in the names of the SSN files that can be downloaded. The job is initiated with the **"Create SSN"** button (black arrow).

The progress of the job is available on the "**Previous Jobs**" panel: "**PENDING**" when the job is received, "**RUNNING**" while the job is running, and a link to the results page when the job is finished. Also, an e-mail is sent to the user when the job is started; a second e-mail is sent when the job is finished that includes a link to the results page.

"SSNs Created from this Dataset" tab. If the initial dataset was used previously to generate SSNs, e.g., with different alignment scores, this tab will be present. The tab provides a list of those SSNs (as links to their "**Download Network Files**" pages; red bracket).



EFI - ENZYME SIMILARITY TOOL

DATASET COMPLETED

Submission Name: **IPR004184_IP74_UniRef90**

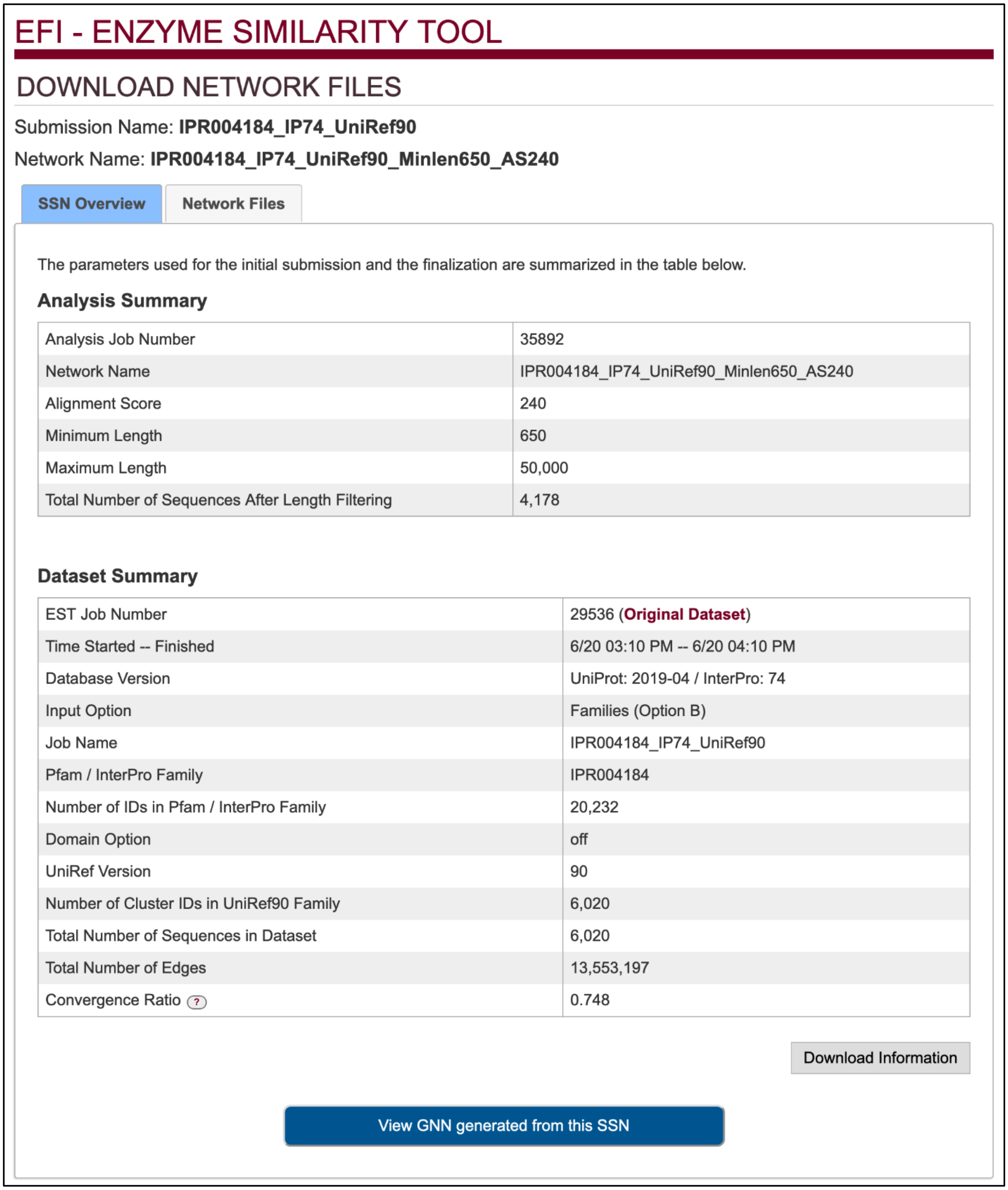
A minimum sequence similarity threshold that specifies the sequence pairs connected by edges is needed to generate the SSN. This threshold also determines the segregation of proteins into clusters. The threshold is applied to the edges in the SSN using the alignment score, an edge node attribute that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

A list of the SSNs generated from this initial dataset.

- **IPR004184_IP74_UniRef90_Minlen650_AS120 AS=120 Min=650 (Analysis ID=35890)**
- **IPR004184_IP74_UniRef90_Minlen650_AS185 AS=185 Min=650 (Analysis ID=35891)**
- **IPR004184_IP74_UniRef90_Minlen650_AS240 AS=240 Min=650 (Analysis ID=35892)**
- **IPR004184_IP74_UniRef90_Minlen650_AS320 AS=320 Min=650 (Analysis ID=35893)**

F. EFI-EST “Download Network Files” page: Downloading SSN files

When the SSN is completed, SSN files (xgmml format; uncompressed or zipped) for viewing with Cytoscape are available for download from this page.



EFI - ENZYME SIMILARITY TOOL

DOWNLOAD NETWORK FILES

Submission Name: **IPR004184_IP74_UniRef90**

Network Name: **IPR004184_IP74_UniRef90_Minlen650_AS240**

The parameters used for the initial submission and the finalization are summarized in the table below.

Analysis Summary

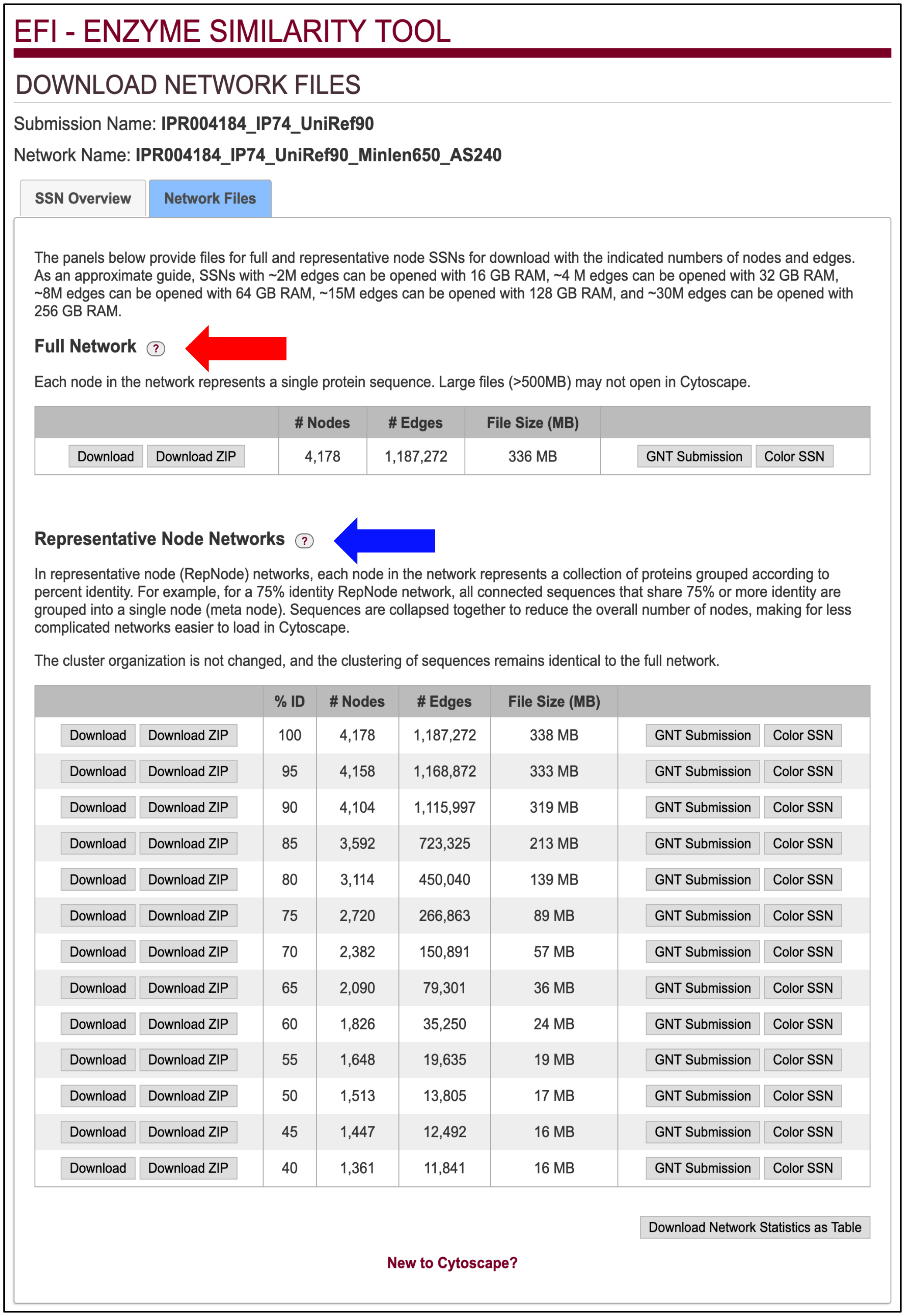
Analysis Job Number	35892
Network Name	IPR004184_IP74_UniRef90_Minlen650_AS240
Alignment Score	240
Minimum Length	650
Maximum Length	50,000
Total Number of Sequences After Length Filtering	4,178

Dataset Summary

EST Job Number	29536 (**Original Dataset**)
Time Started -- Finished	6/20 03:10 PM -- 6/20 04:10 PM
Database Version	UniProt: 2019-04 / InterPro: 74
Input Option	Families (Option B)
Job Name	IPR004184_IP74_UniRef90
Pfam / InterPro Family	IPR004184
Number of IDs in Pfam / InterPro Family	20,232
Domain Option	off
UniRef Version	90
Number of Cluster IDs in UniRef90 Family	6,020
Total Number of Sequences in Dataset	6,020
Total Number of Edges	13,553,197
Convergence Ratio ?	0.748

“SSN Overview” tab. This tab provides 1) the “**Analysis Summary**” table (red arrow) that summarizes the parameters used for “**SSN Finalization**” (*vide supra*) and the total number of sequences in the SSN (after length filtering); and 2) the “**Dataset Summary**” table (blue arrow) that was provided on the “**Dataset Completed**” page. Both can be downloaded as a text file for future reference.

“Network Files” tab. This tab provides access to the full and representative node SSNs, along with the numbers of nodes and edges in each SSN as well as the uncompressed xgmml file size; these statistics can be downloaded as a text file for future reference.



EFI - ENZYME SIMILARITY TOOL

DOWNLOAD NETWORK FILES

Submission Name: **IPR004184_IP74_UniRef90**

Network Name: **IPR004184_IP74_UniRef90_Minlen650_AS240**

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, ~4 M edges can be opened with 32 GB RAM, ~8M edges can be opened with 64 GB RAM, ~15M edges can be opened with 128 GB RAM, and ~30M edges can be opened with 256 GB RAM.

Full Network ?

Each node in the network represents a single protein sequence. Large files (>500MB) may not open in Cytoscape.

	# Nodes	# Edges	File Size (MB)	
Download Download ZIP	4,178	1,187,272	336 MB	GNT Submission Color SSN

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.

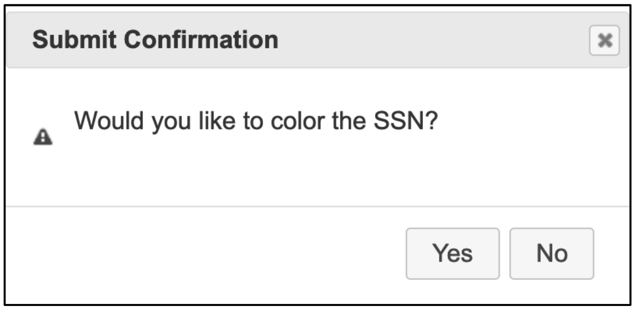
The cluster organization is not changed, and the clustering of sequences remains identical to the full network.

	% ID	# Nodes	# Edges	File Size (MB)	
Download Download ZIP	100	4,178	1,187,272	338 MB	GNT Submission Color SSN
Download Download ZIP	95	4,158	1,168,872	333 MB	GNT Submission Color SSN
Download Download ZIP	90	4,104	1,115,997	319 MB	GNT Submission Color SSN
Download Download ZIP	85	3,592	723,325	213 MB	GNT Submission Color SSN
Download Download ZIP	80	3,114	450,040	139 MB	GNT Submission Color SSN
Download Download ZIP	75	2,720	266,863	89 MB	GNT Submission Color SSN
Download Download ZIP	70	2,382	150,891	57 MB	GNT Submission Color SSN
Download Download ZIP	65	2,090	79,301	36 MB	GNT Submission Color SSN
Download Download ZIP	60	1,826	35,250	24 MB	GNT Submission Color SSN
Download Download ZIP	55	1,648	19,635	19 MB	GNT Submission Color SSN
Download Download ZIP	50	1,513	13,805	17 MB	GNT Submission Color SSN
Download Download ZIP	45	1,447	12,492	16 MB	GNT Submission Color SSN
Download Download ZIP	40	1,361	11,841	16 MB	GNT Submission Color SSN

New to Cytoscape?

As an approximate guide, SSNs with ~2M edges can be opened with 16 GB RAM, ~4 M edges can be opened with 32 GB RAM, ~8M edges can be opened with 64 GB RAM, ~15M edges can be opened with 128 GB RAM, and ~30M edges can be opened with 256 GB RAM.

The "**Full Network**" panel (red arrow) provides access to the full SSN (all UniProt, UniRef50 cluster, or UniRef90 cluster IDs, depending on the database used for generating the SSN). The SSN can be downloaded as an uncompressed ("**Download**" button) or zipped ("**Download ZIP**" button) xgmml file. The file can be transferred directly to EFI-GNT ("**GNT Submission**" button) and to the Color SSNs utility ("**Color SSN**" button). If the transfer to the Color SSNs utility is selected, the user will be asked to verify the submission:



Submit Confirmation

Would you like to color the SSN?

The "**Representative Node Networks**" panel provides access to a set of thirteen representative node (rep node) SSNs is generated in which UniProt accession/UniRef cluster IDs that share from 40 to 100% sequence identity, in 5% increments, are clustered in the same metanode. Although rep node SSNs have decreased resolution relative to full SSNs, their xgmml files are smaller so at least one should be viewable with Cytoscape. The SSNs can be downloaded as uncompressed ("**Download**" button) or zipped ("**Download ZIP**" button) xgmml files. The files can be transferred directly to EFI-GNT ("**GNT Submission**" button) and to the Color SSNs utility ("**Color SSN**" button).

G. Further Refinement of the Minimum Alignment Score Threshold

After the initial SSN is visualized and analyzed using Cytoscape, the minimum alignment score threshold can be refined to achieve isofunctional clusters by selecting a value that separates experimentally validated functions, e.g., as described in the main text, from SwissProt (a node attribute in the SSNs provided by EFI-EST) and/or functions obtained by a survey of the literature and/or unpublished data. If the initial SSN was generated with a minimum alignment score threshold that corresponds to 35 to 40% pairwise identity, it is likely that SSN will have multiple functions in the same clusters (under-fractionated).

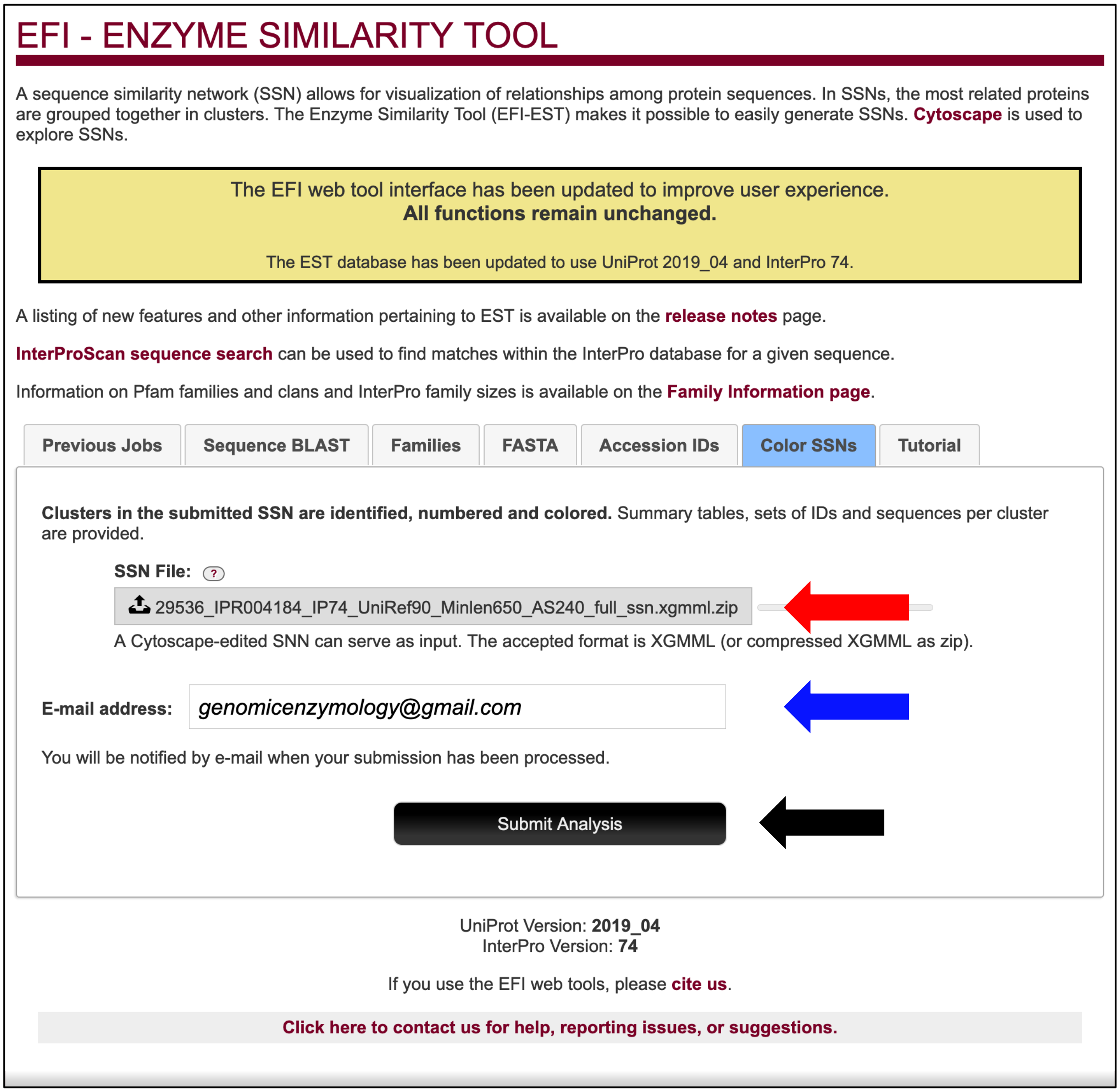
The Color SSNs utility (Section II.H) provides a FASTA file for each SSN cluster that can be used for generating multiple sequence alignments so that residue conservation can be assessed using algorithms such as ClustalOmega (https://www.ebi.ac.uk/Tools/msa/clustalo/) or MUSCLE (https://www.ebi.ac.uk/Tools/msa/muscle/), thereby providing orthogonal information about conserved sequence/function. And, the SSN can be used as the input for EFI-GNT that allows local genome context to be assessed via GNNs that summarize co-occurrence frequencies of protein families that occur within a ±N orf window of the sequences in each SSN cluster and individual genome neighborhood diagrams (GNDs) for the sequences in each SSN cluster, again providing additional evidence about isofunctionality.

Cytoscape can be used to filter/remove edges to generate SSNs with clusters that share greater levels of sequence identity. For large SSNs, it is more convenient to use the "**Dataset Finalization**" tab on the "**Dataset Completed**" page to filter/remove edges with a larger "**Alignment Score Threshold**" because edge removal with Cytoscape can be slow for large SSNs.

Because the analysis step of EFI-EST is fast, we usually generate a series of SSNs at increasing alignment score thresholds and visually inspect these to identify an appropriate alignment score for segregating the SSN into isofunctional clusters.

It is also important to realize that segregation of orthologues across a functionally diverse superfamily with a single alignment score may not be possible because sequence boundaries between functions (substrate specificity and/or reaction mechanism) often do not diverge uniformly as sequence similarity decreases. The user can use Cytoscape to generate daughter networks for functionally heterogeneous clusters that allow more detailed, focused analyses without compromising the integrity of the entire SSN. Daughter networks are created by selecting/highlighting the clusters and then clicking the "New Network from Selection (all edges)" button at the top of the Cytoscape window; larger alignment score thresholds can be applied only to the cluster in the daughter network to further segregate the nodes into multiple clusters that can be assessed for isofunctionality using SwissProt-curated/literature functions and/or genome context.

H. EFI-EST Home Page, "Color SSNs" tab: Utility for identifying and coloring clusters in an SSN



EFI - ENZYME SIMILARITY TOOL

A sequence similarity network (SSN) allows for visualization of relationships among protein sequences. In SSNs, the most related proteins are grouped together in clusters. The Enzyme Similarity Tool (EFI-EST) makes it possible to easily generate SSNs. **Cytoscape** is used to explore SSNs.

The EFI web tool interface has been updated to improve user experience.
All functions remain unchanged.

The EST database has been updated to use UniProt 2019_04 and InterPro 74.

A listing of new features and other information pertaining to EST is available on the **release notes** page.

InterProScan sequence search can be used to find matches within the InterPro database for a given sequence.

Information on Pfam families and clans and InterPro family sizes is available on the **Family Information page**.

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

SSN File: 29536_IPR004184_IP74_UniRef90_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: *genomicenzymology@gmail.com*

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

UniProt Version: **2019_04**
InterPro Version: **74**

If you use the EFI web tools, please **cite us**.

Click here to contact us for help, reporting issues, or suggestions.

The Color SSNs utility assigns a unique color and number (in order of decreasing numbers of sequences in the cluster) to each cluster and a unique number to each singleton. The colored SSN may be useful for more easily describing clusters of interest and/or preparing figures for publications and slides. Also, the cluster and singleton numbers are required by EFI-CGFP for quantitating metagenome abundance (*vide infra*).

SSN File Upload. The user uploads ("**SSN File**"; red arrow) the xgmml file for an SSN (UniProt or UniRef). The SSN can be generated for either full length sequences or domains; it also can be modified with Cytoscape and exported to an xgmml file. The SSN also can be directly transferred from the "**Download Network Files**" page (*vide infra*).

A previously colored SSN that has been edited in Cytoscape (edge-filtered with a larger alignment score or a daughter with selected clusters) also can be used as input; the cluster numbers and colors will be reassigned based on cluster sizes in the edited SSN. Colored SSNs generated by EFI-GNT or modified by the marker identification and marker abundance quantification steps of EFI-CGFP also can be used as input.

On the "**Previous Jobs**" panel and the "**Job History**" page, Color SSN jobs are designated "Color SSN".

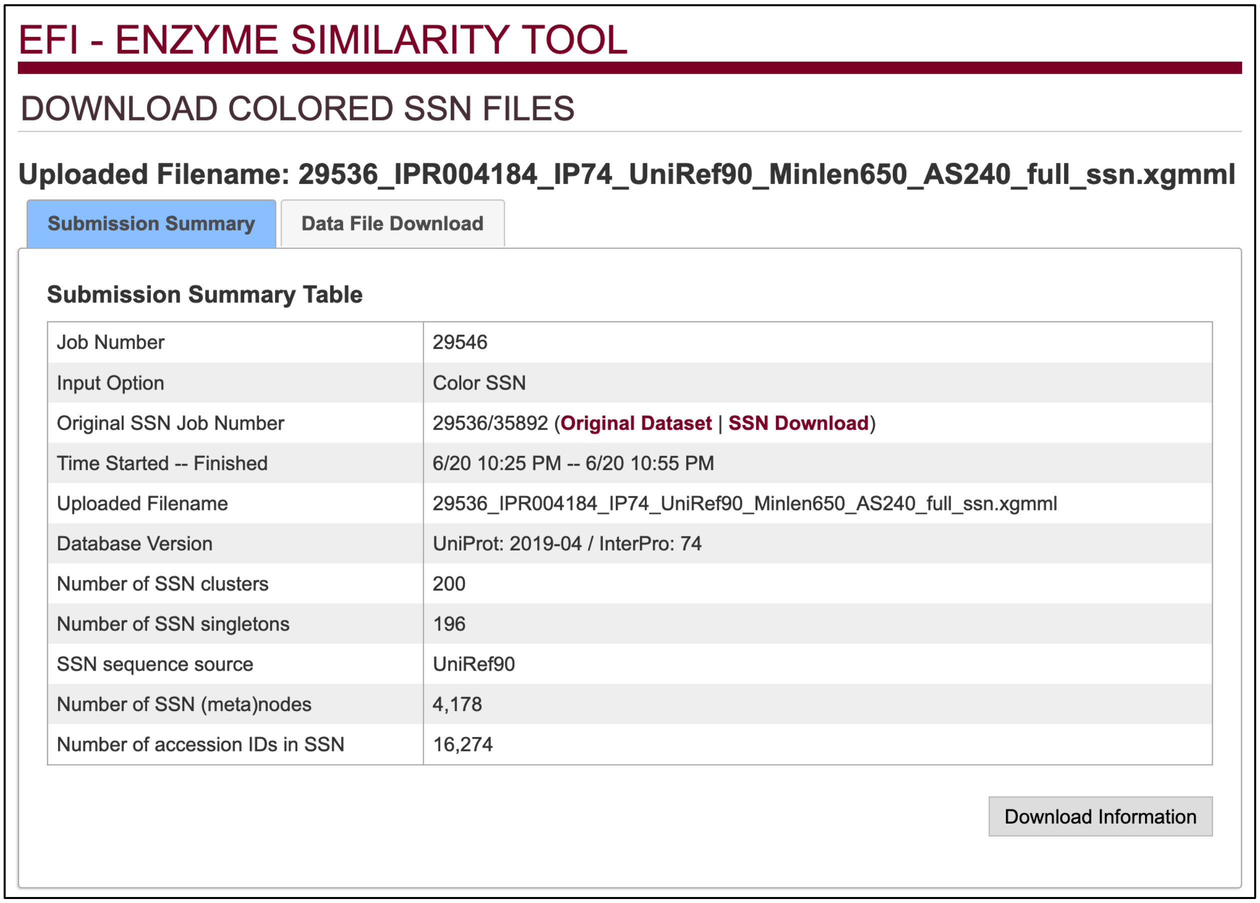
An "**E-mail address**" (blue arrow) is required. The job is initiated with the "**Submit Analysis**" button (black arrow).

The progress of the job is available on the "**Previous Jobs**" panel: "**PENDING**" when the job is received, "**RUNNING**" while the job is running, and a link to the results page when the job is finished. Also, an e-mail is sent to the user when the job is started; a second e-mail is sent when the job is finished that includes a link to the results page.

I. “Download Colored SSN Files” page: Colored SSN

When the Color SSN job is completed, the user is directed to the “**Download Colored SSN Files**” page.

“Submission Summary” tab. This tab provides a table summarizing the job parameters, including the original SNN job number with links to the “**Dataset Completed**” and “**Download Network Files**” pages for generating the SSN. The table also provides the numbers of clusters, singletons, and (meta)nodes in the SSN.



EFI - ENZYME SIMILARITY TOOL

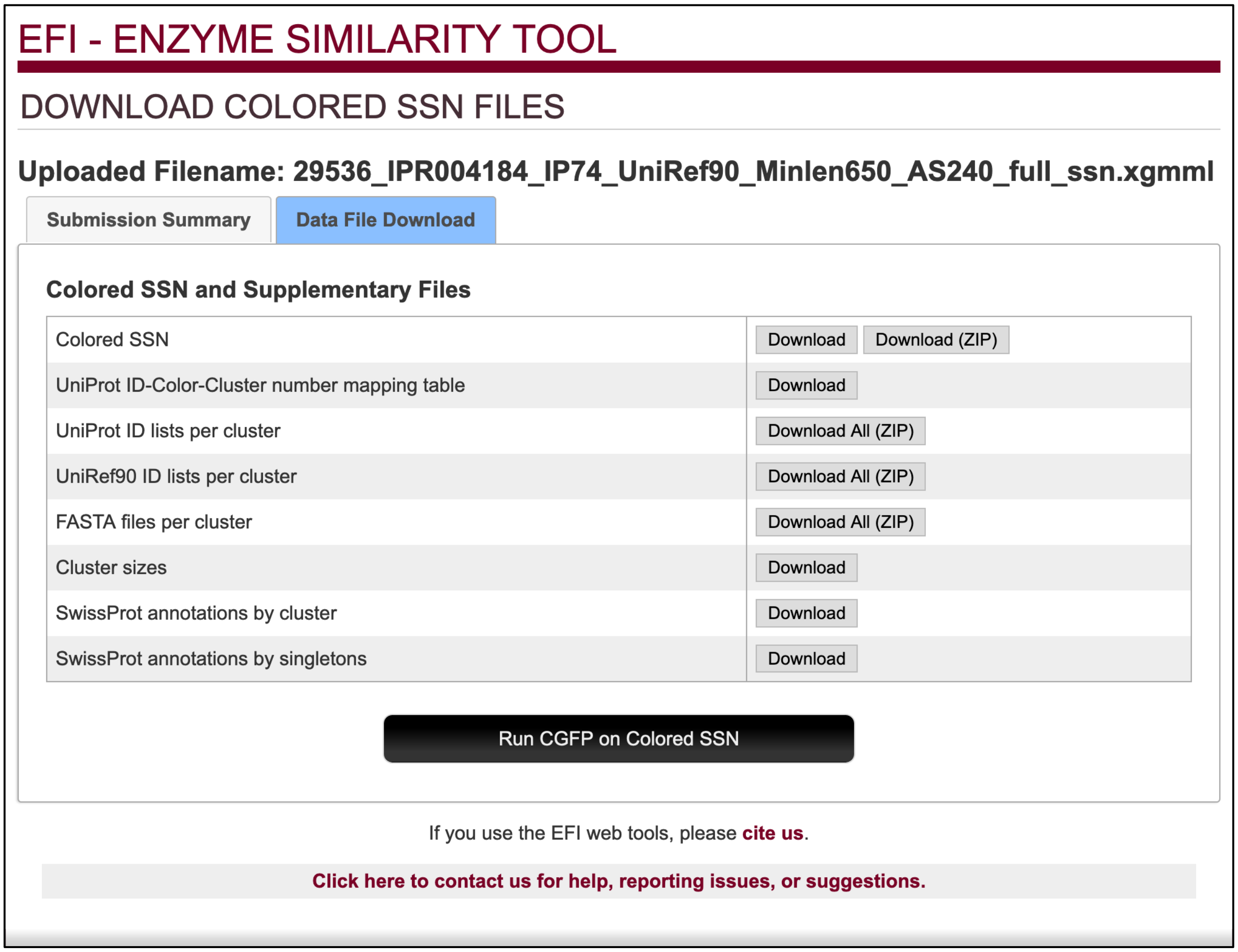
DOWNLOAD COLORED SSN FILES

Uploaded Filename: 29536_IPR004184_IP74_UniRef90_Minlen650_AS240_full_ssn.xgmml

Submission Summary Table

Job Number	29546
Input Option	Color SSN
Original SSN Job Number	29536/35892 (**Original Dataset** | **SSN Download**)
Time Started -- Finished	6/20 10:25 PM -- 6/20 10:55 PM
Uploaded Filename	29536_IPR004184_IP74_UniRef90_Minlen650_AS240_full_ssn.xgmml
Database Version	UniProt: 2019-04 / InterPro: 74
Number of SSN clusters	200
Number of SSN singletons	196
SSN sequence source	UniRef90
Number of SSN (meta)nodes	4,178
Number of accession IDs in SSN	16,274

“Data File Download” tab. This tab provides several files for download:



EFI - ENZYME SIMILARITY TOOL

DOWNLOAD COLORED SSN FILES

Uploaded Filename: 29536_IPR004184_IP74_UniRef90_Minlen650_AS240_full_ssn.xgmml

Colored SSN and Supplementary Files

Colored SSN	Download Download (ZIP)
UniProt ID-Color-Cluster number mapping table	Download
UniProt ID lists per cluster	Download All (ZIP)
UniRef90 ID lists per cluster	Download All (ZIP)
FASTA files per cluster	Download All (ZIP)
Cluster sizes	Download
SwissProt annotations by cluster	Download
SwissProt annotations by singletons	Download

If you use the EFI web tools, please cite us.

Click here to contact us for help, reporting issues, or suggestions.

1) “**Colored SSN**”. The input SSN is modified to include a unique color and number for each cluster (in order of decreasing numbers of sequences in the cluster) and a unique number for each singleton in the input SSN. The cluster and singleton numbers are required for SSNs that are used as the input for EFI-CGFP. Four node attributes are added to the input SSN: “Cluster Number”, “Cluster Sequence Count”, “node.fillColor” (hexadecimal descriptor for the unique color), and “Singleton Number”. The node attributes in the color SSN generated by the Color SSNs utility are included in Table S1.

2) “**UniProt ID-Color-Cluster number mapping table**”. This tab-delimited text file is used with the BridgeDB Cytoscape app (http://apps.cytoscape.org/apps/bridgedb) to color nodes in SSNs containing the same UniProt IDs, e.g., segregated into clusters with different alignments scores and/or daughter SSNs that contain a subset of the sequences/clusters. Coloring SSNs with the same UniProt IDs but clustered with different alignment scores is useful to assess how clusters segregate as the alignment score is increased or aggregate as the alignment score is decreased. If the input SSN was generated using domains, a mapping table also is provided with the domain IDs associated with cluster colors and numbers.

3) “**Cluster ID lists per cluster**”, “**UniRef90 ID lists per cluster**”, and “**UniRef50 ID lists per cluster**”. The Color SSNs utility provides folders of files with lists of the UniProt and UniRef90 IDs in each cluster generated using UniRef90 clusters and UniProt, UniRef90, and UniRef50 IDs in each cluster generated using UniRef50 clusters; these can be used with Option D of EFI-EST to generate higher resolution SSNs for selected clusters (*vide supra*).

4) “**FASTA files per cluster**”. The Color SSNs utility provides a folder with files containing the FASTA sequences in each SSN cluster is available; these files can be used as input for multiple sequence alignment (MSA) algorithms, e.g., CLUSTALW or MUSCLE.

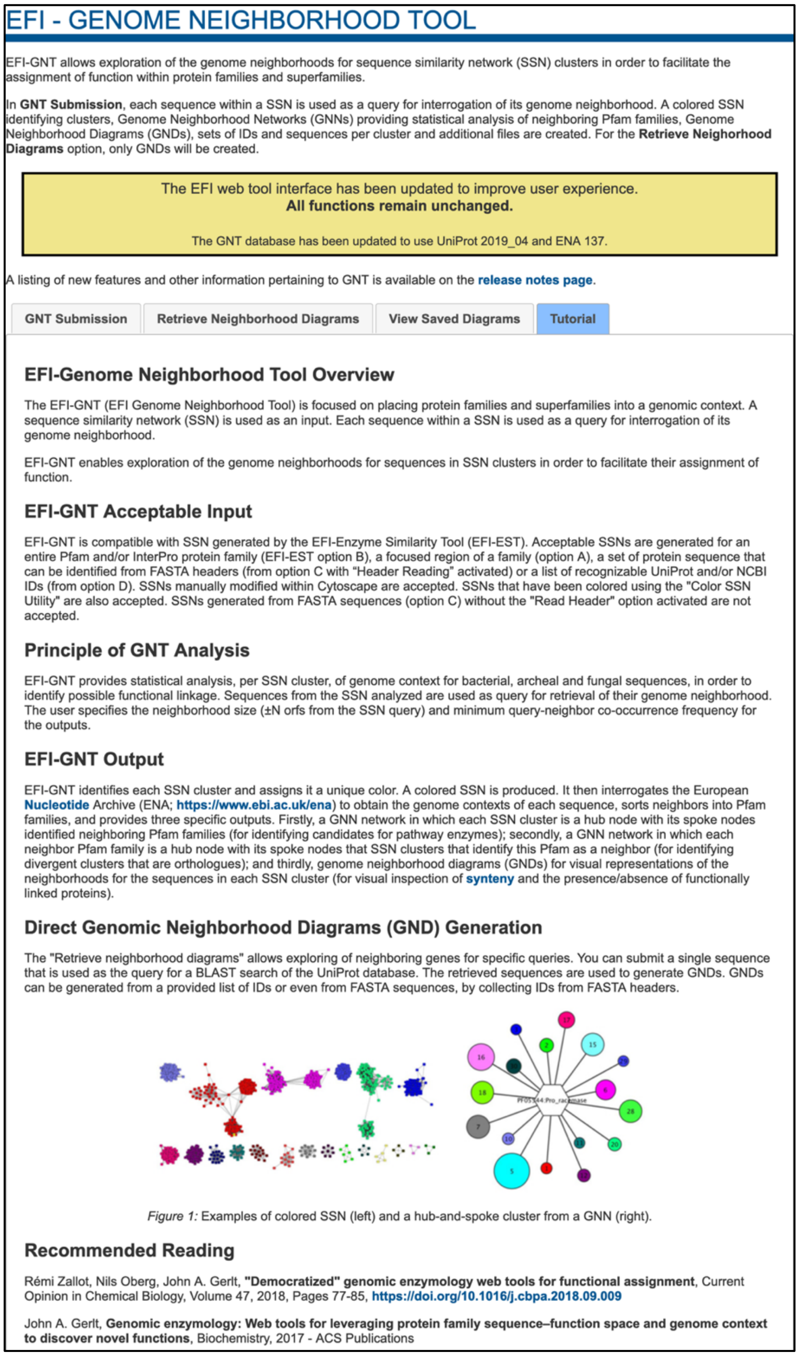
5) “**Cluster sizes**”, “**SwissProt annotations by cluster**”, and “**SwissProt annotations by singletons**”. The Color SSNs utility provides a text file with the number of UniProt IDs in each SSN cluster generated with UniProt IDs, the number of UniProt IDs and UniRef90 cluster IDs in each SSN cluster generated with UniRef90 cluster IDs, and the number

of UniProt IDs, UniRef90 cluster IDs, and UniRef50 cluster IDs in each SSN cluster generated with UniRef50 cluster IDs. It also provides the accession IDs for sequences with SwissProt annotations (separate files for clusters and singletons).

Transfer of Colored SSN to EFI-CGFP. The colored SSN can be directly transferred to EFI-CGFP using the "**Run CGFP on Colored SSN**" button. However, we recommend analyzing/editing the SSN (using, perhaps, the GNNs/GNDs from EFI-GNT) to ensure that the SSN clusters are isofunctional before proceeding with CGFP.

III. EFI-GNT (https://efi.igb.illinois.edu/efi-gnt/)

This section describes EFI-GNT, including selection of parameters for generating genome neighborhood networks (GNNs) and genome neighborhood diagrams (GNDs), and the files available for download.



EFI - GENOME NEIGHBORHOOD TOOL

EFI-GNT allows exploration of the genome neighborhoods for sequence similarity network (SSN) clusters in order to facilitate the assignment of function within protein families and superfamilies.

In **GNT Submission**, each sequence within a SSN is used as a query for interrogation of its genome neighborhood. A colored SSN identifying clusters, Genome Neighborhood Networks (GNNs) providing statistical analysis of neighboring Pfam families, Genome Neighborhood Diagrams (GNDs), sets of IDs and sequences per cluster and additional files are created. For the **Retrieve Neighorhood Diagrams** option, only GNDs will be created.

The EFI web tool interface has been updated to improve user experience.
All functions remain unchanged.

The GNT database has been updated to use UniProt 2019_04 and ENA 137.

A listing of new features and other information pertaining to GNT is available on the **release notes page**.

EFI-Genome Neighborhood Tool Overview

The EFI-GNT (EFI Genome Neighborhood Tool) is focused on placing protein families and superfamilies into a genomic context. A sequence similarity network (SSN) is used as an input. Each sequence within a SSN is used as a query for interrogation of its genome neighborhood.

EFI-GNT enables exploration of the genome neighborhoods for sequences in SSN clusters in order to facilitate their assignment of function.

EFI-GNT Acceptable Input

EFI-GNT is compatible with SSN generated by the EFI-Enzyme Similarity Tool (EFI-EST). Acceptable SSNs are generated for an entire Pfam and/or InterPro protein family (EFI-EST option B), a focused region of a family (option A), a set of protein sequence that can be identified from FASTA headers (from option C with "Header Reading" activated) or a list of recognizable UniProt and/or NCBI IDs (from option D). SSNs manually modified within Cytoscape are accepted. SSNs that have been colored using the "Color SSN Utility" are also accepted. SSNs generated from FASTA sequences (option C) without the "Read Header" option activated are not accepted.

Principle of GNT Analysis

EFI-GNT provides statistical analysis, per SSN cluster, of genome context for bacterial, archeal and fungal sequences, in order to identify possible functional linkage. Sequences from the SSN analyzed are used as query for retrieval of their genome neighborhood. The user specifies the neighborhood size (±N orfs from the SSN query) and minimum query-neighbor co-occurrence frequency for the outputs.

EFI-GNT Output

EFI-GNT identifies each SSN cluster and assigns it a unique color. A colored SSN is produced. It then interrogates the European **Nucleotide** Archive (ENA; **https://www.ebi.ac.uk/ena**) to obtain the genome contexts of each sequence, sorts neighbors into Pfam families, and provides three specific outputs. Firstly, a GNN network in which each SSN cluster is a hub node with its spoke nodes identified neighboring Pfam families (for identifying candidates for pathway enzymes); secondly, a GNN network in which each neighbor Pfam family is a hub node with its spoke nodes that SSN clusters that identify this Pfam as a neighbor (for identifying divergent clusters that are orthologues); and thirdly, genome neighborhood diagrams (GNDs) for visual representations of the neighborhoods for the sequences in each SSN cluster (for visual inspection of **synteny** and the presence/absence of functionally linked proteins).

Direct Genomic Neighborhood Diagrams (GND) Generation

The "Retrieve neighborhood diagrams" allows exploring of neighboring genes for specific queries. You can submit a single sequence that is used as the query for a BLAST search of the UniProt database. The retrieved sequences are used to generate GNDs. GNDs can be generated from a provided list of IDs or even from FASTA sequences, by collecting IDs from FASTA headers.

Figure 1: Examples of colored SSN (left) and a hub-and-spoke cluster from a GNN (right).

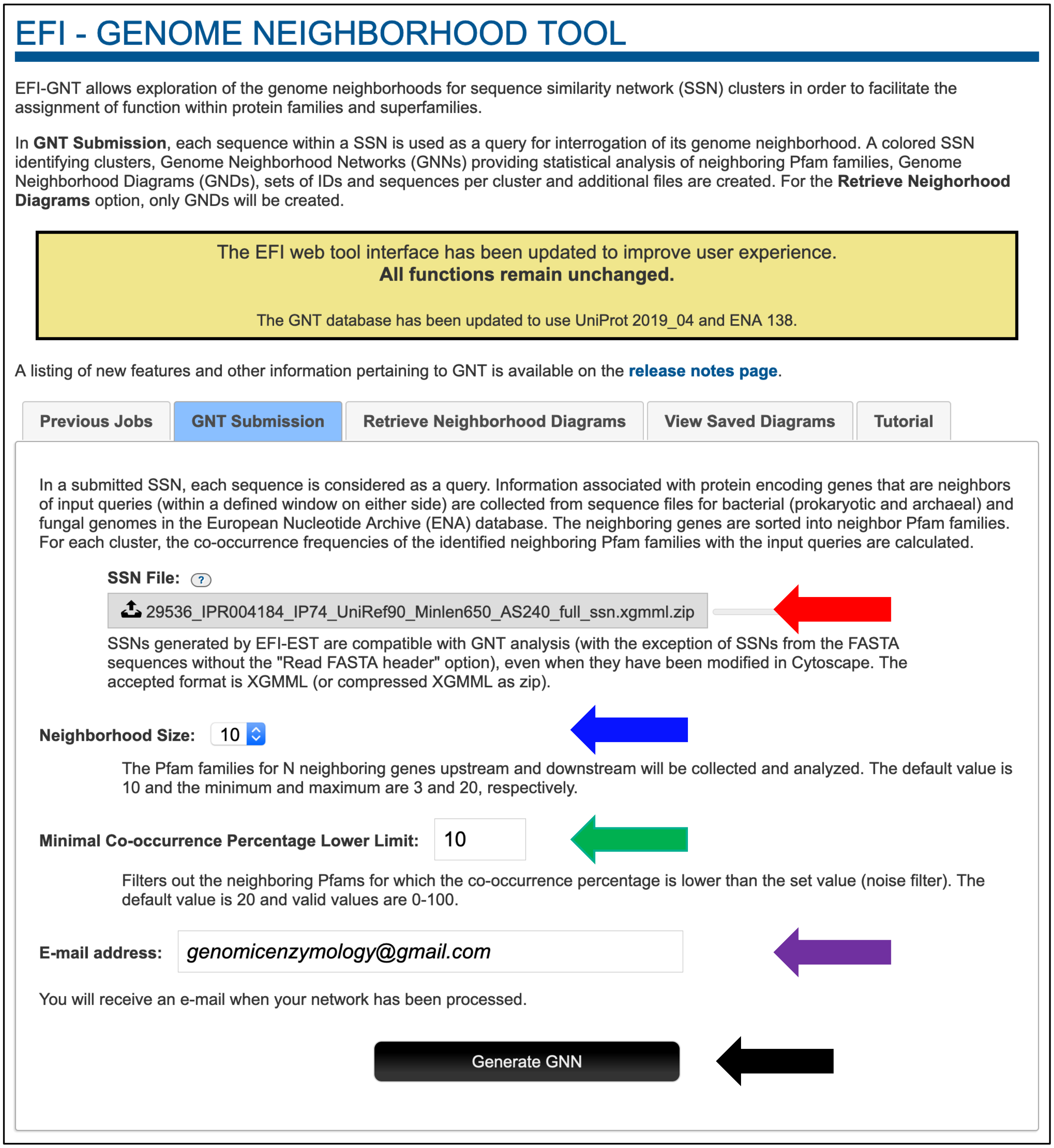
Recommended Reading

Rémi Zallot, Nils Oberg, John A. Gerlt, **"Democratized" genomic enzymology web tools for functional assignment**, Current Opinion in Chemical Biology, Volume 47, 2018, Pages 77-85, **https://doi.org/10.1016/j.cbpa.2018.09.009**

John A. Gerlt, **Genomic enzymology: Web tools for leveraging protein family sequence–function space and genome context to discover novel functions**, Biochemistry, 2017 - ACS Publications

Genome Database for Generating GNNs and GNDs. Nucleic acid (genome) sequences are downloaded from the ENA database (https://www.ebi.ac.uk/ena) with each release (approximately every three months; releases of the UniProt, InterPro, and ENA databases are not coordinated); the downloaded files include the STD (assembled sequences), CON (scaffolds), and WGS (genomic contig) datasets for PRO (prokaryotic/archaeal), ENV (environmental), and FUN (fungal) taxonomic divisions. These are used to populate a local database that includes information about the genome neighborhoods for all UniProt IDs, including start/stop locations for the encoding genes and relative directions of transcription that are used to generate GNDs as well as the identities of the neighbor Pfam and InterPro families. The complete genome neighborhoods collected by EFI-GNT (± 20 orfs) are not available for all UniProt IDs—only the coding sequence is deposited for some proteins; for others, a smaller neighborhood size is determined by the position of the gene on the sequence contig. Eukaryotic sequences are not included in the database.

A. EFI-GNT Home Page, "GNT Submission" tab: Generating GNNs and GNDs from SSNs



EFI - GENOME NEIGHBORHOOD TOOL

EFI-GNT allows exploration of the genome neighborhoods for sequence similarity network (SSN) clusters in order to facilitate the assignment of function within protein families and superfamilies.

In **GNT Submission**, each sequence within a SSN is used as a query for interrogation of its genome neighborhood. A colored SSN identifying clusters, Genome Neighborhood Networks (GNNs) providing statistical analysis of neighboring Pfam families, Genome Neighborhood Diagrams (GNDs), sets of IDs and sequences per cluster and additional files are created. For the **Retrieve Neighorhood Diagrams** option, only GNDs will be created.

The EFI web tool interface has been updated to improve user experience.
All functions remain unchanged.

The GNT database has been updated to use UniProt 2019_04 and ENA 138.

A listing of new features and other information pertaining to GNT is available on the **release notes page**.

In a submitted SSN, each sequence is considered as a query. Information associated with protein encoding genes that are neighbors of input queries (within a defined window on either side) are collected from sequence files for bacterial (prokaryotic and archaeal) and fungal genomes in the European Nucleotide Archive (ENA) database. The neighboring genes are sorted into neighbor Pfam families. For each cluster, the co-occurrence frequencies of the identified neighboring Pfam families with the input queries are calculated.

SSN File: ?

29536_IPR004184_IP74_UniRef90_Minlen650_AS240_full_ssn.xgmml.zip

SSNs generated by EFI-EST are compatible with GNT analysis (with the exception of SSNs from the FASTA sequences without the "Read FASTA header" option), even when they have been modified in Cytoscape. The accepted format is XGMML (or compressed XGMML as zip).

Neighborhood Size: 10

The Pfam families for N neighboring genes upstream and downstream will be collected and analyzed. The default value is 10 and the minimum and maximum are 3 and 20, respectively.

Minimal Co-occurrence Percentage Lower Limit: 10

Filters out the neighboring Pfams for which the co-occurrence percentage is lower than the set value (noise filter). The default value is 20 and valid values are 0-100.

E-mail address: *genomicenzymology@gmail.com*

You will receive an e-mail when your network has been processed.

SSN Entry. The user uploads the in SSN xgmml file (uncompressed or zipped) ("**SSN File**"; red arrow). SSNs generated by EFI-EST can be modified in Cytoscape, e.g., filtered to remove edges with larger alignment scores or daughter networks that contain a subset of the clusters in the starting SSN. Alternatively, an SSN can be transferred from the "**Download Network Files**" page of EFI-EST to EFI-GNT (*vide supra*). The input SSN can be generated for either full length sequences or domains.

EFI-GNT can retrieve genome neighborhood information only for sequences in SSNs that have UniProt accession IDs because these are required to access the genome sequence entries in the ENA database; therefore, SSNs generated with Option C without the "**Read FASTA header**" option are not compatible with EFI-GNT.

The user specifies a "**Neighborhood Size**" (± N orfs from the SSN query sequences, with N ≤20; default 10; blue arrow) and a "**Minimal Co-occurrence Percentage Lower Limit**" (from 0 to 100%; default 20%; green arrow) between the SSN cluster query sequences and their genome neighbors for the statistical analyses that are used to generate the GNN files as well as the GNDs. The values of the "**Neighborhood Size**" and "**Minimal Co-occurrence Percentage Lower Limit**" are used to calculate the co-occurrence frequencies in the GNNs (*vide infra*); the value of the "**Neighborhood Size**" determines the initial display presented in the GND viewer, although the genome neighborhood window can be changed with the viewer (from ±1 to ±20 orfs).

The SSN file name is used as the job name on the "**Previous Jobs**" tab and the "**Job History**" page; the jobs are designated "**GNN**".

An "**E-mail address**" (magenta arrow) is required. The job is initiated with the "**Generate GNN**" button (black arrow).

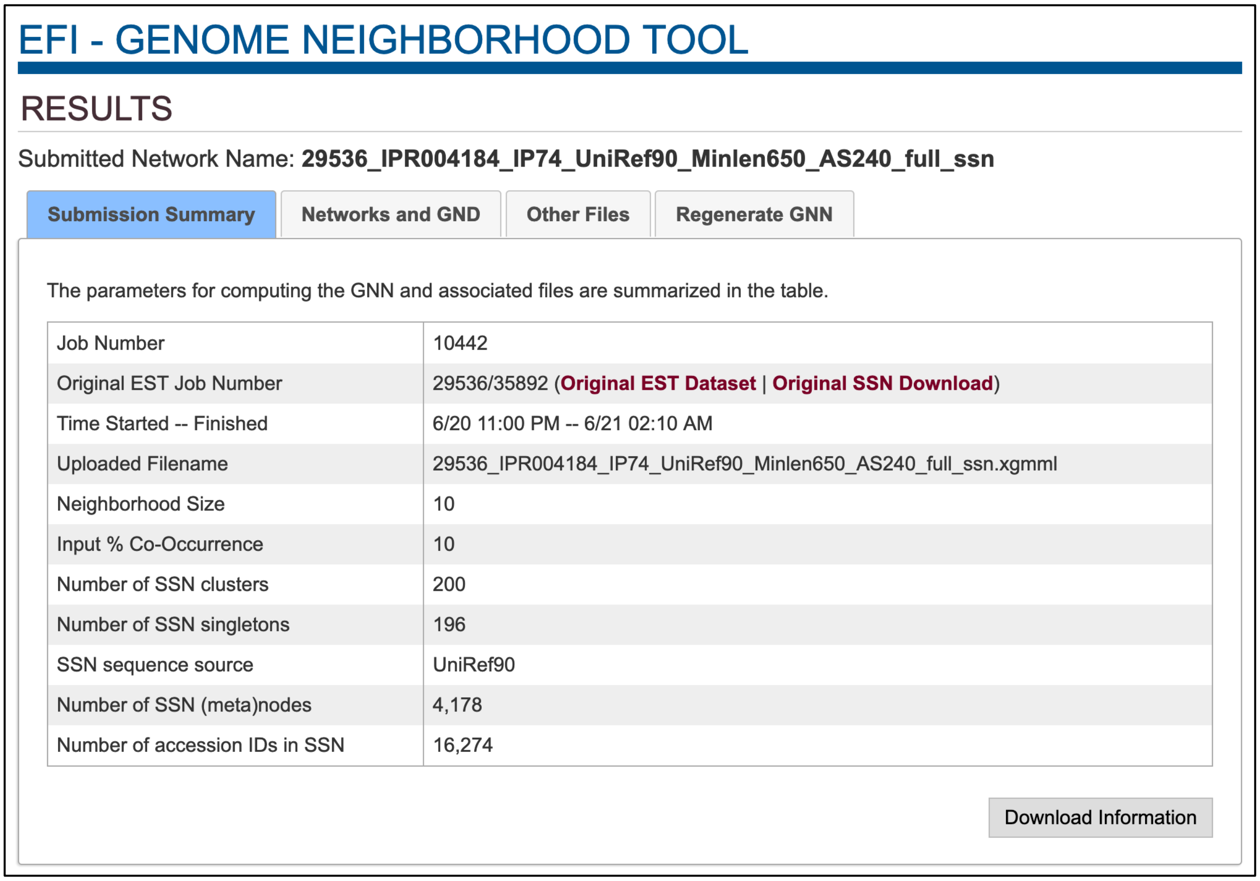
The progress of the job is available on the "**Previous Jobs**" panel: "**PENDING**" when the job is received, "**RUNNING**" while the job is running, and a link to the results page when the job is finished. Also, an e-mail is sent to the user when the job is started; a second e-mail is sent when the job is finished that includes a link to the results page.

Irrespective of the user-specified value for the "**Neighborhood Size**", EFI-GNT collects all genome neighbors within ±20 orfs of the query sequences (or the maximum number possible in each direction if the query UniProt ID is located near/at the end of a contig) and associates these with Pfam families and InterPro domains, families, and homologous superfamilies. EFI-GNT then 1) filters the neighbors using the user-specified genome neighborhood window and minimum co-occurrence frequency and 2) analyzes the genome context data using these values.

Approximately 25% of Pfam families are "DUFs", Domains of Unknown Function; ~5% of the entries in UniProt are members of these families. In addition, 23% of the sequences in UniProt (currently, >35M sequences) are not associated with an InterPro domain/family (including Pfam). Thus, it is likely multiple genome neighbors identified by EFI-GNT will not be associated with a Pfam family; these are included in the "None" family in the GNNs.

B. EFI-GNT "Results" Page, Downloading Colored SSN and GNNs, and Accessing GND Viewer

When these analyses are completed, the user is directed to the EFI-GNT "**Results**" page.



EFI - GENOME NEIGHBORHOOD TOOL

RESULTS

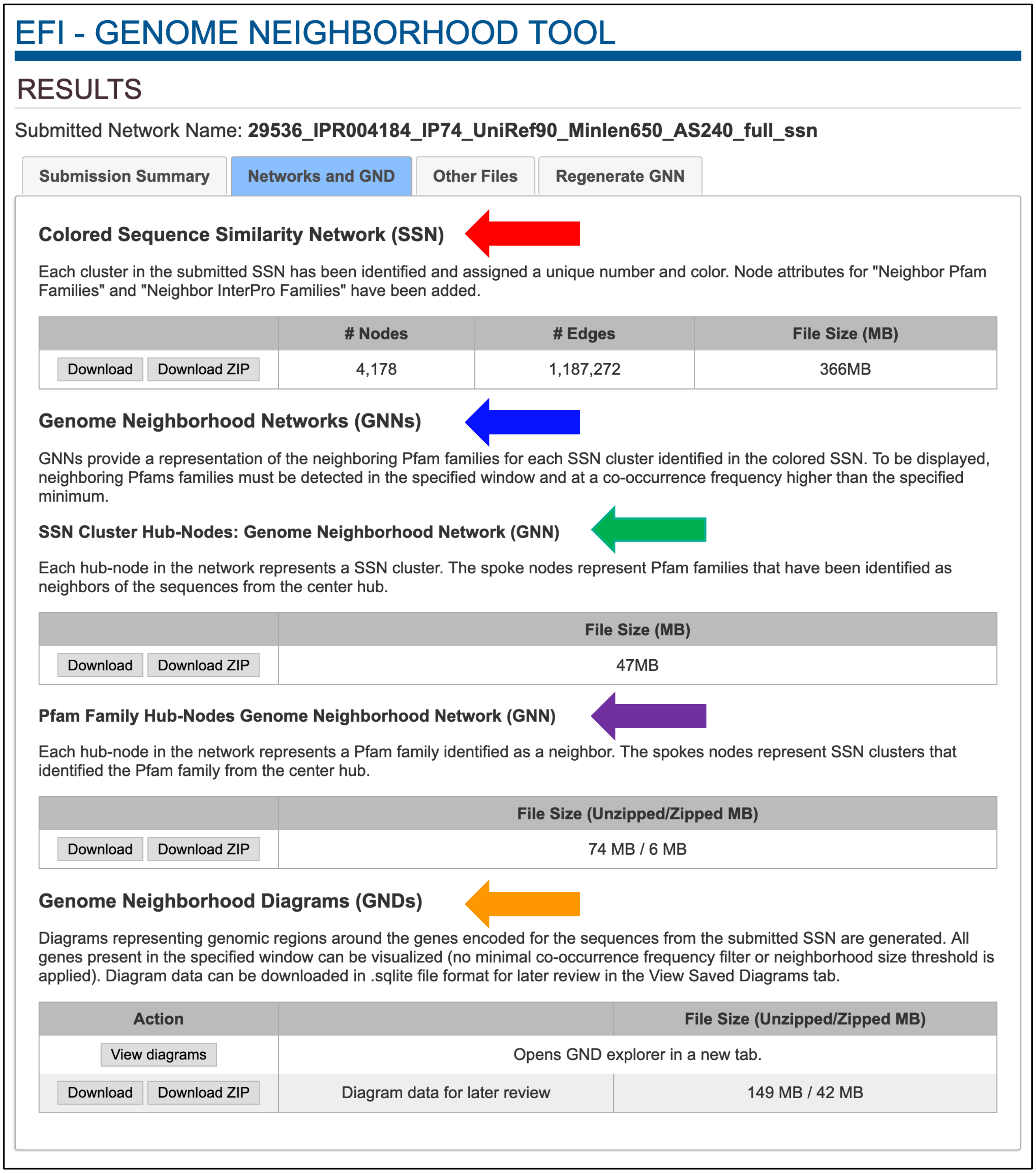
Submitted Network Name: **29536_IPR004184_IP74_UniRef90_Minlen650_AS240_full_ssn**

The parameters for computing the GNN and associated files are summarized in the table.

Job Number	10442
Original EST Job Number	29536/35892 (**Original EST Dataset** | **Original SSN Download**)
Time Started -- Finished	6/20 11:00 PM -- 6/21 02:10 AM
Uploaded Filename	29536_IPR004184_IP74_UniRef90_Minlen650_AS240_full_ssn.xgmml
Neighborhood Size	10
Input % Co-Occurrence	10
Number of SSN clusters	200
Number of SSN singletons	196
SSN sequence source	UniRef90
Number of SSN (meta)nodes	4,178
Number of accession IDs in SSN	16,274

"Submission Summary" tab. This tab provides a table that details the job parameters, including links to the original EFI-EST "**Dataset Completed**" and "**Download Network Files**" pages, the number of accession IDs and (meta)nodes in the SSN, and the numbers of SSN clusters and singletons; the table can be downloaded as a text file for future reference.

"Networks and GND" tab. This tab provides access to the Colored SSN generated by EFI-GNT, two complementary GNNs, and the GNDs:



RESULTS

Submitted Network Name: **29536_IPR004184_IP74_UniRef90_Minlen650_AS240_full_ssn**

Colored Sequence Similarity Network (SSN)

Each cluster in the submitted SSN has been identified and assigned a unique number and color. Node attributes for "Neighbor Pfam Families" and "Neighbor InterPro Families" have been added.

	# Nodes	# Edges	File Size (MB)
Download Download ZIP	4,178	1,187,272	366MB

Genome Neighborhood Networks (GNNs)

GNNs provide a representation of the neighboring Pfam families for each SSN cluster identified in the colored SSN. To be displayed, neighboring Pfams families must be detected in the specified window and at a co-occurrence frequency higher than the specified minimum.

SSN Cluster Hub-Nodes: Genome Neighborhood Network (GNN)

Each hub-node in the network represents a SSN cluster. The spoke nodes represent Pfam families that have been identified as neighbors of the sequences from the center hub.

	File Size (MB)
Download Download ZIP	47MB

Pfam Family Hub-Nodes Genome Neighborhood Network (GNN)

Each hub-node in the network represents a Pfam family identified as a neighbor. The spokes nodes represent SSN clusters that identified the Pfam family from the center hub.

	File Size (Unzipped/Zipped MB)
Download Download ZIP	74 MB / 6 MB

Genome Neighborhood Diagrams (GNDs)

Diagrams representing genomic regions around the genes encoded for the sequences from the submitted SSN are generated. All genes present in the specified window can be visualized (no minimal co-occurrence frequency filter or neighborhood size threshold is applied). Diagram data can be downloaded in .sqlite file format for later review in the View Saved Diagrams tab.

Action		File Size (Unzipped/Zipped MB)
View diagrams	Opens GND explorer in a new tab.	
Download Download ZIP	Diagram data for later review	149 MB / 42 MB

1) “**Colored Sequence Similarity Network (SSN)**” (red arrow). A colored SSN similar to that produced by the Color SSNs utility (i.e., node attributes for cluster numbers, sizes, and colors as well as singleton numbers; domain-delimited “names” and “shared names” for the nodes if the a domain SSN was used as input), with additional node attributes indicating the protein is present in the local ENA database (“Present in ENA Database?”: “true” for bacterial, archaeal, and fungal proteins; otherwise “false”), the presence/absence of neighbors in the ENA genome sequence files (“Genome Neighbors in ENA Database?”: “true” or “false” for bacterial, archaeal, and fungal proteins; otherwise “n/a”), the identity of the ENA file (“ENA Database Genome ID”), and the identities of the neighbor Pfam families (“Neighbor Pfam Families”) and the neighbor InterPro families (“Neighbor InterPro Families”) identified by the bacterial, archaeal, and fungal proteins in each (meta)node in the SSN [using all UniProt IDs, not just UniRef clusters (if UniRef is used to generate the SSN)] within the user-specified genome neighborhood window size *but independent of the user-specified query-neighbor co-occurrence frequency*. The node attributes in the colored SSN generated by EFI-GNT are listed in Table S3.

Using the Select panel of Cytoscape, the “Neighbor Pfam Families” and “Neighbor InterPro Families” node attributes allow the user to associate specific neighbor Pfam and InterPro families with nodes/clusters in the SSN, complementing the information contained in the GNNs (next two paragraphs). Our experience is that selecting nodes/clusters based on the identities of the neighbor families in the colored SSN can be more effective/efficient for identifying functionally linked proteins than interpreting the hub/spoke-node GNNs.

Two GNNs are available for download from the "**Genome Neighborhood Networks (GNNs)**" panel (blue arrow):

2) "**SSN Cluster Hub-Nodes Genome Neighborhood Network (GNN)**" (green arrow). A GNN containing clusters with SSN cluster-hub nodes (the queries for identifying neighbors) and Pfam family-spoke nodes (the neighbor Pfam families that were identified in the user-specified neighborhood window) is available for download. The neighbors used to construct these spoke nodes are chosen using the user-specified co-occurrence frequency and neighborhood orf window size. The clusters facilitate identification of neighbors (associated with Pfam families) that may be functionally linked to the query, e.g., participate in the same metabolic pathway. In favorable situations, it may be possible to identify the types, if not the identities, of the reactions in the pathway by inspection of the Pfam families identified in the spoke nodes. The node attributes in this GNN file are listed in Table S4; these include SSN query-genome neighborhood co-occurrence frequencies and query-neighbor distances (in orfs).

3) "**Pfam Family Hub-Nodes Genome Neighborhood Network (GNN)**" (magenta arrow). A GNN containing clusters with Pfam family-hub nodes (the neighbor Pfam families that were identified in the user-specified neighborhood window) and SSN cluster-spoke nodes (the SSN queries for identifying neighbors) is available for download. The information presented in this GNN is the "same" as in the previous GNN, except the identities of the hub and spoke nodes are reversed. These clusters facilitate the identification of potentially orthologous SSN clusters, e.g., the same Pfam family (hub node) is identified by multiple clusters (spoke nodes). In some cases, different SSN clusters may identify paralogues, not orthologues, in the same Pfam family.

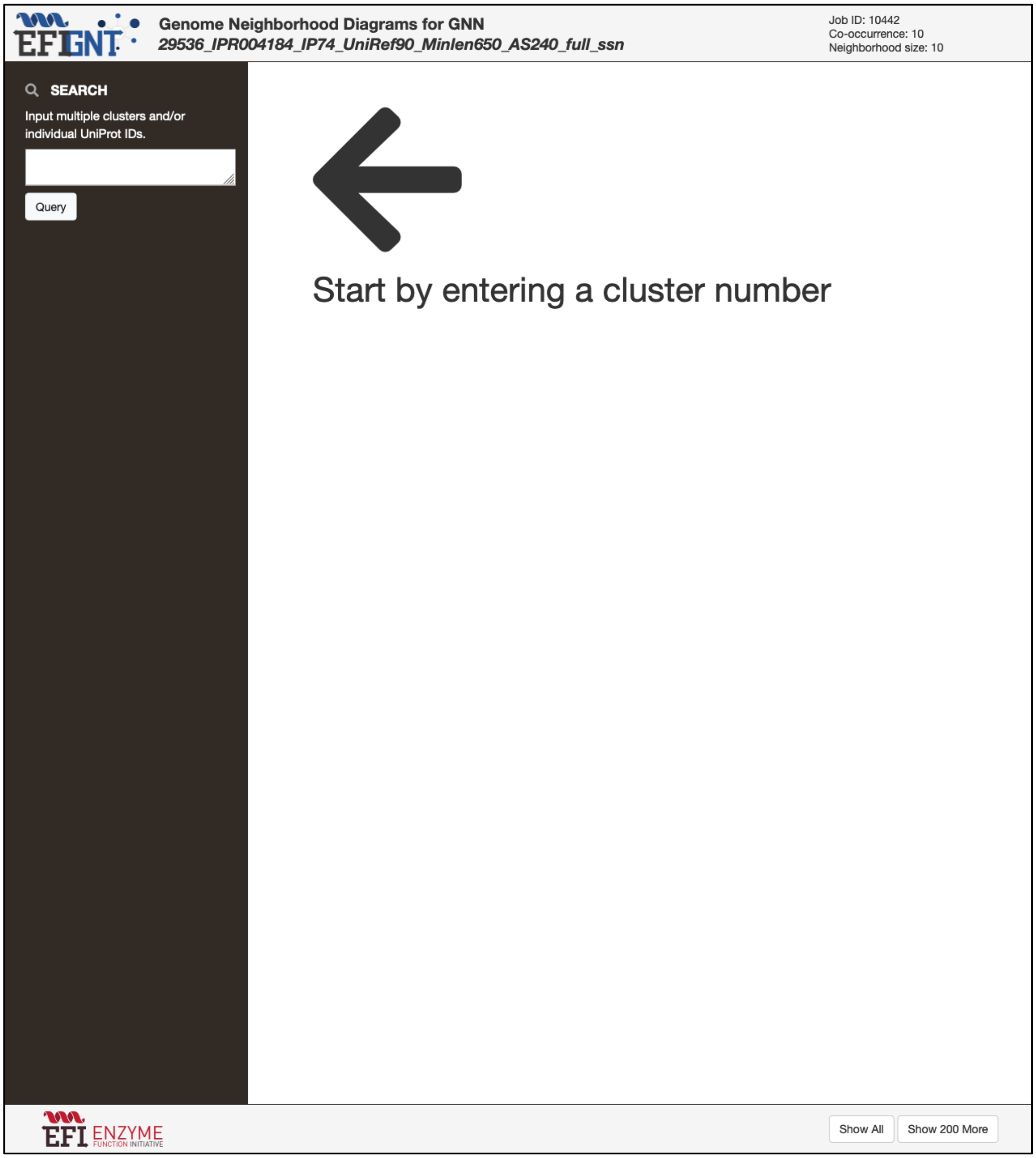
The node attributes in this GNN file are listed in Table S5; these also include SSN query-genome neighborhood co-occurrence frequencies and query-neighbor distances (in orfs).

4) “**Genome Neighborhood Diagrams (GNDs)**” (orange arrow). This panel provides access to the “GND Explorer” that displays the individual genome neighborhoods identified in the ENA database that were used to generate the GNNs; these are designated “genome neighborhood diagrams” (GNDs).

The GNDs generated by EFI-GNT are analogous to those available from the JGI Integrated Microbial Genomes and Microbiomes tool (https://img.jgi.doe.gov/), *with the important difference that EFI-GNT enables input of multiple sequences from “isofunctional” clusters in defined sequence-function space in SSNs.* GNDs allow exploration of local genome context, providing a much richer level of potential functional linkages than is possible using the GNNs with their reports (as node attributes) of SSN query-genome neighbor co-occurrence frequencies and query-neighbor distances (Tables S5 and S6). The GNDs allow the discovery of “rare” species-specific genome contexts that might include all of the genes encoding a pathway (gene clustering/synteny are not conserved across species). “Complete” GNDs (±N orfs, the size of the user-specified N orf window) are not available for all UniProt IDs—only the coding sequence is available for some proteins; for others, the availability/number of neighbors is determined by the position of the protein on the deposited sequence contig.

This panel also allows the user to download the GNDs [as a .sqlite file; either uncompressed (“**Download**” or zipped “**Download ZIP**”)] for saving and later uploading/analysis using the “**View Saved Diagrams**” tab on the EFI-GNT home page.

To view the GNDs with the “GND Explorer”, the user clicks the “**View diagrams**” button in the “**Genome Neighborhood Diagrams (GNDs)**” panel. The “**Genome Neighborhood Diagrams for GNN**” page opens; the user enters one or more SSN cluster numbers in the “**Search**” panel and clicks the “Query” button. (One or more UniProt IDs in the SSN can be entered; cluster numbers and UniProt IDs can be mixed and matched, and the order in which they are displayed is determined based on the input query.)



Genome Neighborhood Diagrams for GNN

29536_IPR004184_IP74_UniRef90_Minlen650_AS240_full_ssn

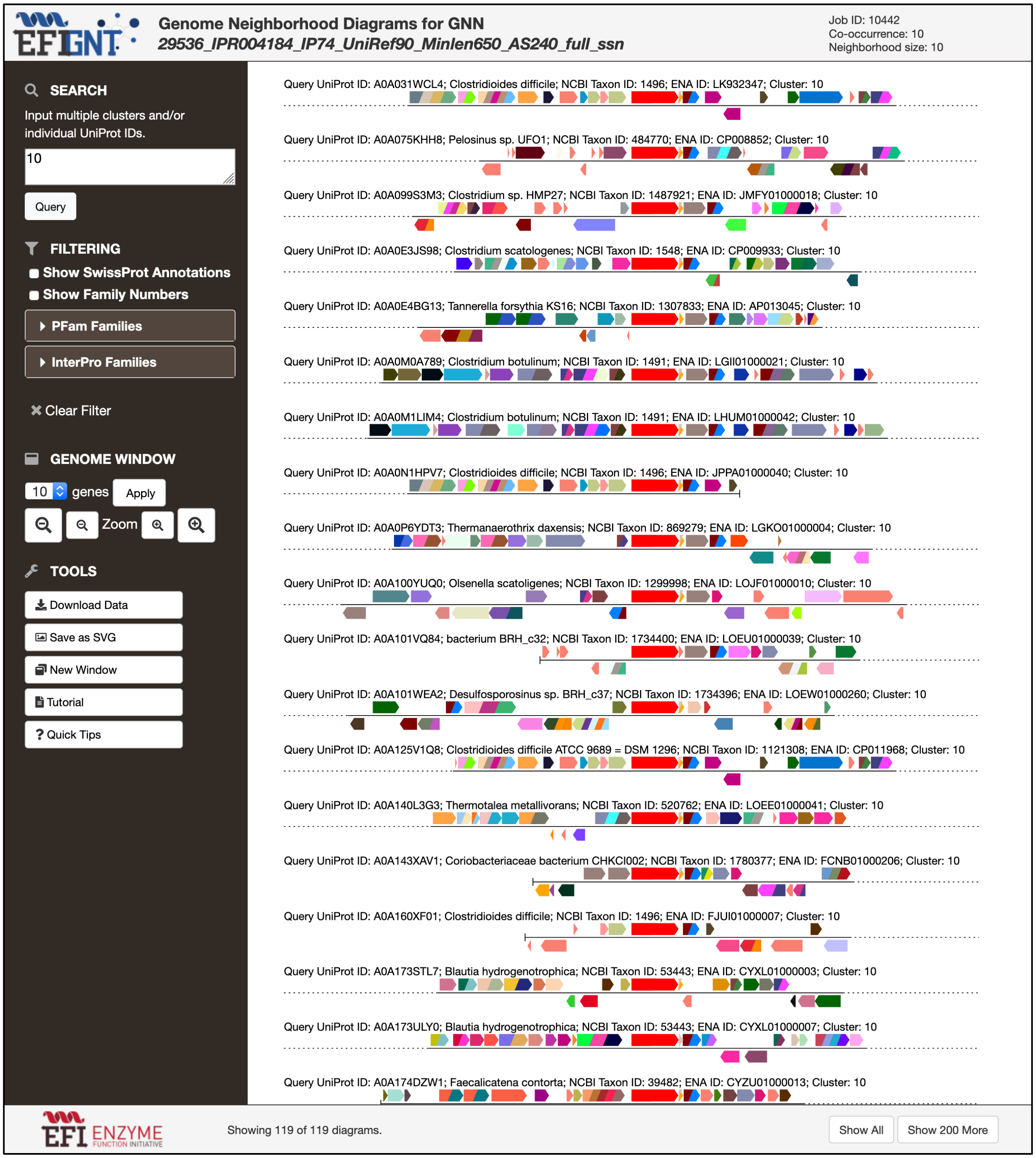
Job ID: 10442
Co-occurrence: 10
Neighborhood size: 10

SEARCH

Input multiple clusters and/or individual UniProt IDs.

Start by entering a cluster number

The GNDs for bacterial, archaeal, and fungal queries in the SSN cluster then are displayed.



Genome Neighborhood Diagrams for GNN
29536_IPR004184_IP74_UniRef90_Minlen650_AS240_full_ssn

Job ID: 10442
Co-occurrence: 10
Neighborhood size: 10

Input multiple clusters and/or individual UniProt IDs.

10

Query UniProt ID: A0A031WCL4; Clostridioides difficile; NCBI Taxon ID: 1496; ENA ID: LK932347; Cluster: 10

Query UniProt ID: A0A075KHH8; Pelosinus sp. UFO1; NCBI Taxon ID: 484770; ENA ID: CP008852; Cluster: 10

Query UniProt ID: A0A099S3M3; Clostridium sp. HMP27; NCBI Taxon ID: 1487921; ENA ID: JMFY01000018; Cluster: 10

Query UniProt ID: A0A0E3JS98; Clostridium scatologenes; NCBI Taxon ID: 1548; ENA ID: CP009933; Cluster: 10

Query UniProt ID: A0A0E4BG13; Tannerella forsythia KS16; NCBI Taxon ID: 1307833; ENA ID: AP013045; Cluster: 10

Query UniProt ID: A0A0M0A789; Clostridium botulinum; NCBI Taxon ID: 1491; ENA ID: LGII01000021; Cluster: 10

Query UniProt ID: A0A0M1LIM4; Clostridium botulinum; NCBI Taxon ID: 1491; ENA ID: LHUM01000042; Cluster: 10

Query UniProt ID: A0A0N1HPV7; Clostridioides difficile; NCBI Taxon ID: 1496; ENA ID: JPPA01000040; Cluster: 10

Query UniProt ID: A0A0P6YDT3; Thermanaerothrix daxensis; NCBI Taxon ID: 869279; ENA ID: LGKO01000004; Cluster: 10

Query UniProt ID: A0A100YUQ0; Olsenella scatoligenes; NCBI Taxon ID: 1299998; ENA ID: LOJF01000010; Cluster: 10

Query UniProt ID: A0A101VQ84; bacterium BRH_c32; NCBI Taxon ID: 1734400; ENA ID: LOEU01000039; Cluster: 10

Query UniProt ID: A0A101WEA2; Desulfosporosinus sp. BRH_c37; NCBI Taxon ID: 1734396; ENA ID: LOEW01000260; Cluster: 10

Query UniProt ID: A0A125V1Q8; Clostridioides difficile ATCC 9689 = DSM 1296; NCBI Taxon ID: 1121308; ENA ID: CP011968; Cluster: 10

Query UniProt ID: A0A140L3G3; Thermotalea metallivorans; NCBI Taxon ID: 520762; ENA ID: LOEE01000041; Cluster: 10

Query UniProt ID: A0A143XAV1; Coriobacteriaceae bacterium CHKCI002; NCBI Taxon ID: 1780377; ENA ID: FCNB01000206; Cluster: 10

Query UniProt ID: A0A160XF01; Clostridioides difficile; NCBI Taxon ID: 1496; ENA ID: FJUI01000007; Cluster: 10

Query UniProt ID: A0A173STL7; Blautia hydrogenotrophica; NCBI Taxon ID: 53443; ENA ID: CYXL01000003; Cluster: 10

Query UniProt ID: A0A173ULY0; Blautia hydrogenotrophica; NCBI Taxon ID: 53443; ENA ID: CYXL01000007; Cluster: 10

Query UniProt ID: A0A174DZW1; Faecalicatena contorta; NCBI Taxon ID: 39482; ENA ID: CYZU01000013; Cluster: 10

Showing 119 of 119 diagrams.

A maximum of 200 GNDs is initially displayed; GNDs for additional queries in the cluster, if any, can be displayed by clicking either the “**Show All**” or “**Show 200 More**” button at the bottom of the page; the progress of the download is displayed. The proteins encoded by the genes in each

GND are displayed as arrows (lengths proportional to the length of the protein, orientations showing the relative directions of transcription). Members of the same Pfam family have the same color so that potential orthologues in different GNDs can be identified; domains in multidomain proteins are colored sequentially with the Pfam family colors for the various domains.

Clicking an arrow displays a window with metadata about the protein: description, UniProt ID, annotation status (TrEMBL or SwissProt), Pfam family(ies), InterPro families(ies)/domain(s), and sequence length. The UniProt ID is a link to the UniProt page for the protein. Clicking the symbol in the upper right-hand corner copies the metadata to the clipboard so that it can be copied/saved in a file.

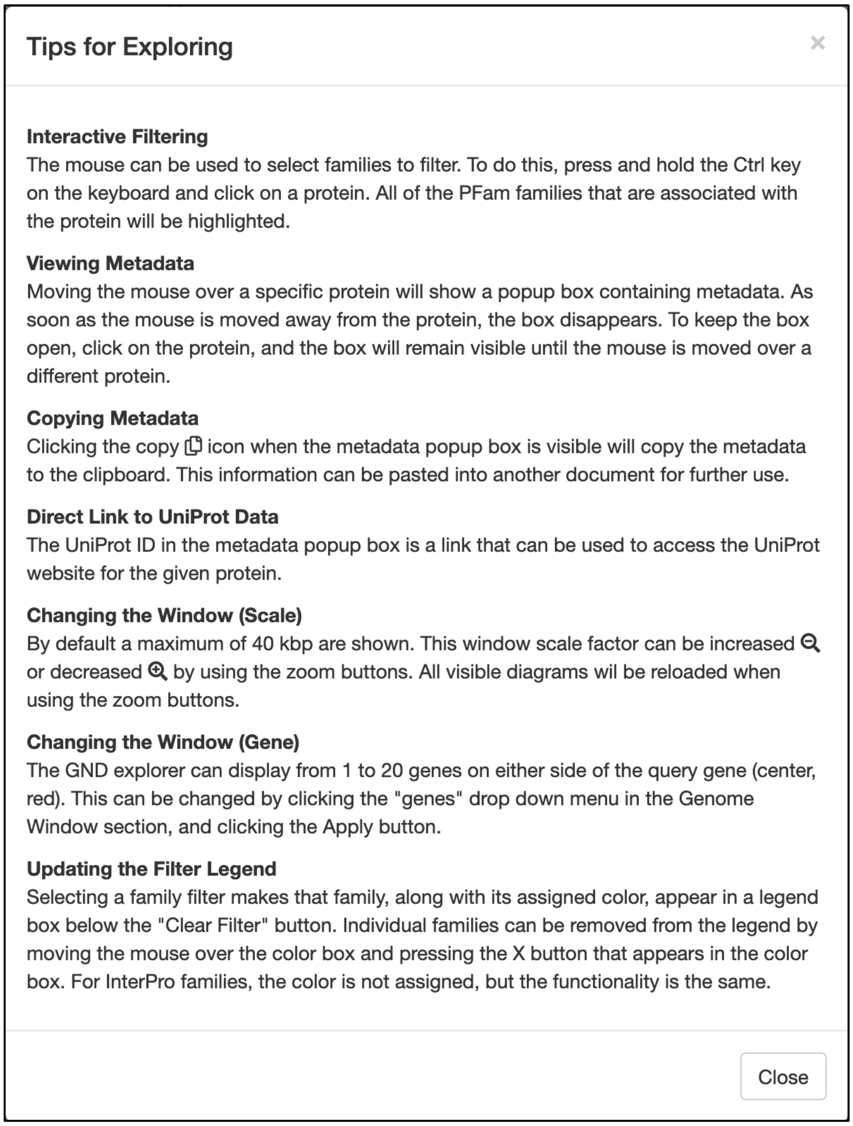
Proteins in the GNDs can be selected by pressing the Control key and clicking on the arrow. The selected protein and all other members in the same Pfam family in all of the displayed GNDs are highlighted by a black border; the colors of the remaining proteins are dimmed. The Pfam families for the selected proteins are displayed (with the Pfam family color) in the legend box below the "**Clear Filter**" button in the left panel. The number of GNDs with selected proteins is displayed in the "**Number of Diagrams Selected**" at the bottom of the viewer; when multiple proteins are selected, the number of diagrams is the number that contains *all* of the selected proteins, enabling the identification of GNDs with the same (functionally linked) proteins.

Proteins can be deselected by pressing the Control key and clicking on the arrow; all proteins that are members of the same Pfam family also are deselected. Members of individual Pfam families can be deselected by clicking the Pfam color in the legend box; the highlighted arrows in the Pfam family will be deselected, unless they are multidomain and have domains that remain in the legend box. All proteins in all families (Pfam and InterPro) can be deselected by clicking the "**Clear Filter**" X.

The left panel provides additional options for exploring and analyzing the GNDs. Proteins with SwissProt-curated annotations are identified by clicking "**Show SwissProt Annotations**". Two accordion panels are available for selecting members of specific Pfam and InterPro families (one accordion can be opened). The families are displayed as lists of short family names (followed by family number) or list of family numbers (followed by short description); the type of list is selected with the "**Show Family Numbers**" click box. When a family is selected, the family numbers/names and Pfam colors are displayed in the legend box (InterPro families are not assigned colors because the GND coloring is based on Pfam domains and their color).

The size of the displayed genome neighborhood window can be varied from ±1 to ±20, irrespective of the window size initially selected to generate the GNNs and display the GNDs. A "Zoom" function is provided to increase/decrease the size of the display.

The "**? Quick Tips**" button at the bottom of the left panel opens a window with descriptions of options for exploring the GNDs:



Tips for Exploring

Interactive Filtering
The mouse can be used to select families to filter. To do this, press and hold the Ctrl key on the keyboard and click on a protein. All of the PFam families that are associated with the protein will be highlighted.

Viewing Metadata
Moving the mouse over a specific protein will show a popup box containing metadata. As soon as the mouse is moved away from the protein, the box disappears. To keep the box open, click on the protein, and the box will remain visible until the mouse is moved over a different protein.

Copying Metadata
Clicking the copy ⧉ icon when the metadata popup box is visible will copy the metadata to the clipboard. This information can be pasted into another document for further use.

Direct Link to UniProt Data
The UniProt ID in the metadata popup box is a link that can be used to access the UniProt website for the given protein.

Changing the Window (Scale)
By default a maximum of 40 kbp are shown. This window scale factor can be increased 🔍 or decreased 🔍 by using the zoom buttons. All visible diagrams wil be reloaded when using the zoom buttons.

Changing the Window (Gene)
The GND explorer can display from 1 to 20 genes on either side of the query gene (center, red). This can be changed by clicking the "genes" drop down menu in the Genome Window section, and clicking the Apply button.

Updating the Filter Legend
Selecting a family filter makes that family, along with its assigned color, appear in a legend box below the "Clear Filter" button. Individual families can be removed from the legend by moving the mouse over the color box and pressing the X button that appears in the color box. For InterPro families, the color is not assigned, but the functionality is the same.

1) "**Interactive Filtering**": The mouse can be used to select families to filter. To do this, press and hold the Ctrl key on the keyboard and click on a protein. All of the members of the Pfam families that are associated with the selected protein will be highlighted. The Pfam families will be displayed in the left panel so that they can be individually deselected (*vide infra*).

2) “**Viewing Metadata**”: Moving the mouse over a specific protein will show a popup box containing metadata about the protein. As soon as the mouse is moved away from the protein, the box disappears. To keep the box open, click on the protein, and the box will remain visible until the mouse is moved over a different protein.

3) “**Copying Metadata**”: Clicking the copy icon when the metadata popup box is visible will copy the metadata to the clipboard. This information can be pasted into another document for further use.

4) “**Link to UniProt Data**”: The UniProt ID in the metadata popup box is a link that can be used to access the UniProt website for the given protein.

5) “**Changing the Window (Scale)**”: By default, a maximum of 40 kbp of genome sequence is shown. This window scale factor can be increased or decreased by using the zoom buttons. All visible diagrams will be reloaded when using the zoom buttons.

6) “**Changing the Window (Gene)**”: The GND Explorer can display from 1 to 20 genes on either side of the query gene (center, red). This can be changed by clicking the "genes" drop down menu in the Genome Window section and clicking the Apply button.

7) “**Updating the Filter Legend**”: Selecting a family filter from the accordions (*vide supra*) or by selecting a protein in a GND (*vide supra*) makes that family, along with its assigned color, appear in a legend box below the "Clear Filter" button. Individual families can be removed

from the legend by moving the mouse over the color box and pressing the X button that appears in the color box. For InterPro families that are selected with the accordion, a color is not assigned, but the functionality is the same.

Other buttons in the left panel provide other capabilities:

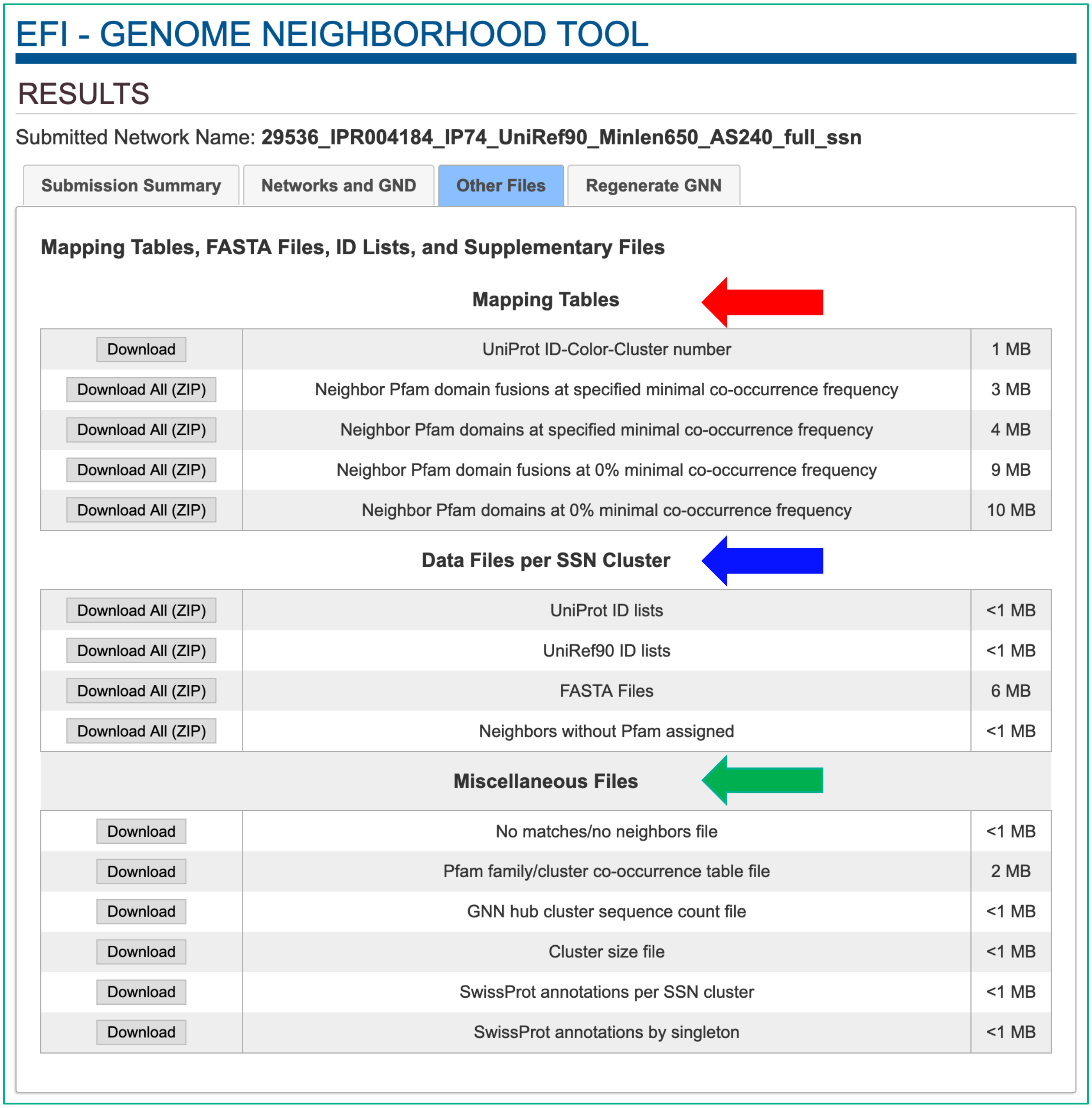
1) "**Download Data**": GNDs can be downloaded and saved for future analyses using the "**View Saved Diagrams**" tool on the EFI-GNT home page.

2) "**Save as SVG"**: The displayed GNDs can be downloaded as an SVG file for manipulation in Adobe Illustrator or the free vector graphics editor Inkscape.

3) "**New Window**": A new window can be opened for display of GNDs.

4) "**Tutorial**": A tutorial is available that provides additional information about viewing and manipulating GNDs.

“**Other Files” tab**. In addition to the GNNs and GNDs, text files are available for download that can facilitate downstream analyses. These files are generated for genome neighbors within the user-specified co-occurrence frequencies and neighborhood windows, unless otherwise specified.



EFI - GENOME NEIGHBORHOOD TOOL

RESULTS

Submitted Network Name: **29536_IPR004184_IP74_UniRef90_Minlen650_AS240_full_ssn**

Mapping Tables, FASTA Files, ID Lists, and Supplementary Files

Mapping Tables

Download	UniProt ID-Color-Cluster number	1 MB
Download All (ZIP)	Neighbor Pfam domain fusions at specified minimal co-occurrence frequency	3 MB
Download All (ZIP)	Neighbor Pfam domains at specified minimal co-occurrence frequency	4 MB
Download All (ZIP)	Neighbor Pfam domain fusions at 0% minimal co-occurrence frequency	9 MB
Download All (ZIP)	Neighbor Pfam domains at 0% minimal co-occurrence frequency	10 MB

Data Files per SSN Cluster

Download All (ZIP)	UniProt ID lists	<1 MB
Download All (ZIP)	UniRef90 ID lists	<1 MB
Download All (ZIP)	FASTA Files	6 MB
Download All (ZIP)	Neighbors without Pfam assigned	<1 MB

Miscellaneous Files

Download	No matches/no neighbors file	<1 MB
Download	Pfam family/cluster co-occurrence table file	2 MB
Download	GNN hub cluster sequence count file	<1 MB
Download	Cluster size file	<1 MB
Download	SwissProt annotations per SSN cluster	<1 MB
Download	SwissProt annotations by singleton	<1 MB

Mapping Tables (red arrow). Several files/folders are available that facilitate mapping of information generated by EFI-GNT to SSNs with the same (or subset) UniProt IDs (mapping table) or provide additional information about genome neighborhoods (incorporated in the GNDs):

1) “**UniProt ID-Color-Cluster number mapping table**” (identical to the mapping table generated by the Color SSNs utility). This tab-delimited text file is used with the BridgeDB Cytoscape app (http://apps.cytoscape.org/apps/bridgedb) to color nodes in SSNs containing the same UniProt IDs, e.g., segregated into clusters with different alignments scores and/or daughter SSNs that contain a subset of the sequences/clusters. Coloring SSNs with the same UniProt IDs but clustered with different alignment scores is useful to assess how clusters segregate as the alignment score is increased or aggregate as the alignment score is decreased. If the input SSN was generated using domains, a mapping table also is provided with the domain IDs associated with cluster colors and numbers.

2) “**Neighbor Pfam domain fusions at specified minimal co-occurrence frequency**”. This folder contains tab-delimited text files for each neighbor Pfam domain fusion architecture that was identified ≥ specified minimal co-occurrence frequency. The file includes the UniProt ID of the query, the UniProt ID of each neighbor, the SSN query cluster number, the SSN query cluster color, the query-neighbor distance, and the query-neighbor directions of transcription (normal, left to right in the genome neighborhood; complement, right to left in the genome neighborhood).

3) “**Neighbor Pfam domains at specified minimal co-occurrence frequency**”. This folder contains tab-delimited text files for each neighbor Pfam domain that was identified ≥ specified minimal co-occurrence frequency. The file includes the UniProt ID of the query, the UniProt ID of each neighbor, the SSN query cluster number, the SSN query cluster color, the query-neighbor distance, and the query-neighbor directions of transcription (normal, left to right in the genome neighborhood; complement, right to left in the genome neighborhood).

4) “**Neighbor Pfam domain fusions at 0% minimal co-occurrence frequency**”. This folder contains tab-delimited text files for each neighbor Pfam domain fusion architecture that was identified. The file includes the UniProt ID of the query, the UniProt ID of each neighbor, the SSN query cluster number, the SSN query cluster color, the query-neighbor distance, and the query-neighbor directions of transcription (normal, left to right in the genome neighborhood; complement, right to left in the genome neighborhood).

5) “**Neighbor Pfam domain fusions at 0% minimal co-occurrence frequency**”. This folder contains tab-delimited text files for each neighbor Pfam domain that was identified. The file includes the UniProt ID of the query, the UniProt ID of each neighbor, the SSN query cluster number, the SSN query cluster color, the query-neighbor distance, and the query-neighbor directions of transcription (normal, left to right in the genome neighborhood; complement, right to left in the genome neighborhood).

“Data Files per SSN Cluster” (identical to files provided by the Color SSNs utility; blue arrow):

1) “**Cluster ID lists per cluster**”, “**UniRef90 ID lists per cluster**”, and “**UniRef50 ID lists per cluster**”. Folders of files with the UniProt IDs in each SSN cluster generated with UniProt IDs, the UniProt IDs and UniRef90 cluster IDs in each SSN cluster generated with UniRef90 cluster IDs, and the UniProt IDs, UniRef90 cluster IDs, and UniRef50 cluster IDs in each SSN cluster generated with UniRef50 cluster IDs; these can be used with Option D of EFI-EST to generate higher resolution SSNs for selected clusters (*vide supra*).

2) “**FASTA files per cluster**”. Folder with files containing the FASTA sequences in each SSN cluster is available; these files can be used as input for multiple sequence alignment (MSA) algorithms, e.g., CLUSTALW or MUSCLE.

3) “**Neighbors without Pfam assigned**”. Folder with files for each cluster containing the UniProt IDs for neighbors not assigned to a Pfam family.

Miscellaneous files (green arrow). Other files are available that provide information about the GNNs:

1) “**No matches/no neighbors file**”. A text file listing the SSN query IDs that did not have matches in the ENA database (no matches) or for which the ENA file did not contain information about genome context (no neighbors).

2) “**Pfam family/cluster co-occurrence table file**”. A tab-delimited text file with the co-occurrence frequencies for each Pfam family in each SSN cluster.

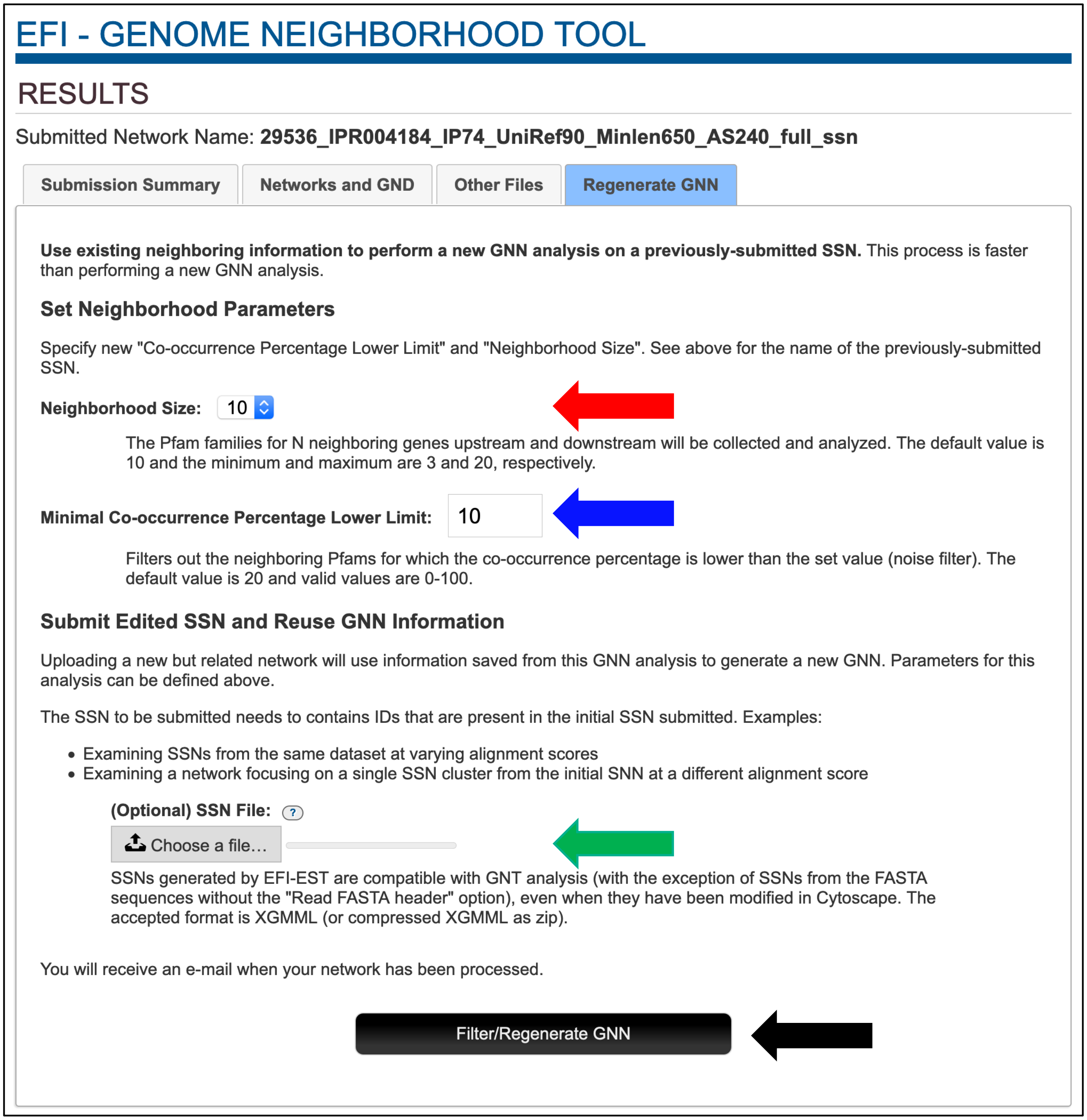
3) “**GNN hub cluster sequence count file**”. A tab-delimited text file with the number of queriable sequences in each cluster and the total number of sequences in each cluster.

4) “**Cluster size file**”. EFI-GNT provides a text file with the number of UniProt IDs in each SSN cluster generated with UniProt IDs, the number of UniProt IDs and UniRef90 cluster IDs in each SSN cluster generated with UniRef90 cluster IDs, and the number of UniProt IDs, UniRef90 cluster IDs, and UniRef50 cluster IDs in each SSN cluster generated with UniRef50 cluster IDs. It also provides the accessions IDs for sequences with SwissProt annotations (separate files for clusters and singletons).

5) “**SwissProt annotations by cluster**”. A tab-delimited text file listing for each cluster UniProt IDs with SwissProt descriptions/annotations.

6) “**SwissProt annotations by singleton**”. A tab-delimited text file listing UniProt IDs with SwissProt descriptions/annotations.

"Regenerate GNN" tab. After the GNNs and GNDs have been generated for an input SSN, the user can use the "**Regenerate GNN**" tab to select different values for the "**Neighborhood Size**" (red arrow) and/or a "**Minimum Co-occurrence Percentage Lower Limit**" (blue arrow) to generate "child" GNNs without recollecting the genome neighborhood information.



EFI - GENOME NEIGHBORHOOD TOOL

RESULTS

Submitted Network Name: **29536_IPR004184_IP74_UniRef90_Minlen650_AS240_full_ssn**

Use existing neighboring information to perform a new GNN analysis on a previously-submitted SSN. This process is faster than performing a new GNN analysis.

Set Neighborhood Parameters

Specify new "Co-occurrence Percentage Lower Limit" and "Neighborhood Size". See above for the name of the previously-submitted SSN.

Neighborhood Size: 10

The Pfam families for N neighboring genes upstream and downstream will be collected and analyzed. The default value is 10 and the minimum and maximum are 3 and 20, respectively.

Minimal Co-occurrence Percentage Lower Limit: 10

Filters out the neighboring Pfams for which the co-occurrence percentage is lower than the set value (noise filter). The default value is 20 and valid values are 0-100.

Submit Edited SSN and Reuse GNN Information

Uploading a new but related network will use information saved from this GNN analysis to generate a new GNN. Parameters for this analysis can be defined above.

The SSN to be submitted needs to contains IDs that are present in the initial SSN submitted. Examples:

- Examining SSNs from the same dataset at varying alignment scores
- Examining a network focusing on a single SSN cluster from the initial SNN at a different alignment score

(Optional) SSN File: ?

SSNs generated by EFI-EST are compatible with GNT analysis (with the exception of SSNs from the FASTA sequences without the "Read FASTA header" option), even when they have been modified in Cytoscape. The accepted format is XGMML (or compressed XGMML as zip).

You will receive an e-mail when your network has been processed.

The user also can upload an SSN [“**(Optional) SSN File**” green arrow] that contains a subset of the sequences in the original SSN or an SSN with the same sequences that was generated with a different alignment score. Filtering/regenerating is much faster than generating the initial GNN because these analyses use the previously collected genome neighborhood data. The job is initiated with the “**Filter/Regenerate GNN**” button (black arrow).

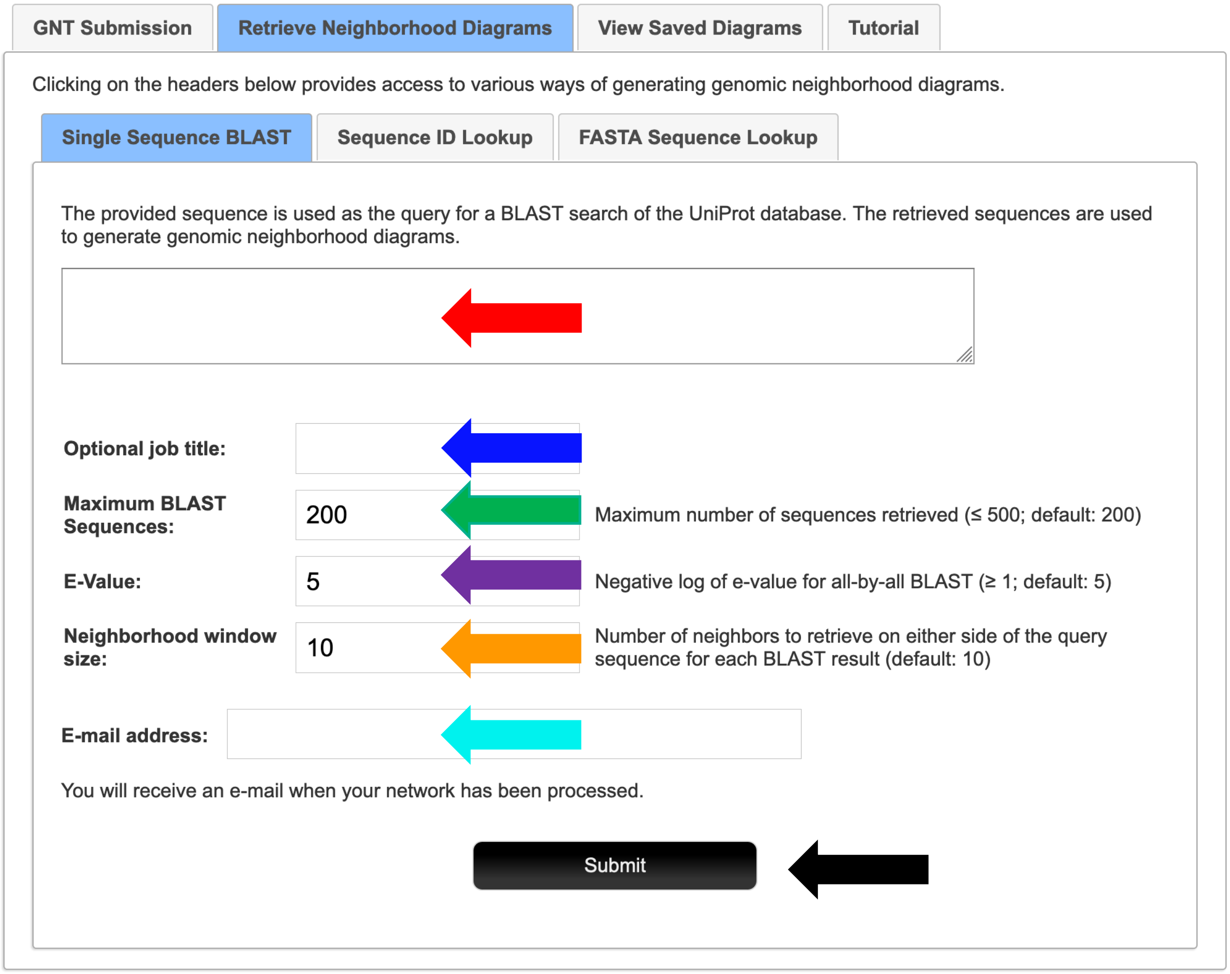
On the “**Previous Jobs**” panel, these jobs are designated “**GNN**”.

The progress of the job is available on the “**Previous Jobs**” panel: “**PENDING**” when the job is received, “**RUNNING**” while the job is running, and a link to the results page when the job is finished. Also, an e-mail is sent to the user when the job is started; a second e-mail is sent when the job is finished that includes a link to the results page.

C. EFI-GNT Home Page, "Retrieve Neighborhood Diagrams" tab: Alternate input methods to generate GNDs

This tab provides access to tools to collect genome neighborhood/context information:

"Single Sequence BLAST" tab: The GNDs for BLAST-identified homologues of an input sequence (entered in the box; red arrow) are retrieved.



Clicking on the headers below provides access to various ways of generating genomic neighborhood diagrams.

The provided sequence is used as the query for a BLAST search of the UniProt database. The retrieved sequences are used to generate genomic neighborhood diagrams.

Optional job title:

Maximum BLAST Sequences: 200 Maximum number of sequences retrieved (≤ 500; default: 200)

E-Value: 5 Negative log of e-value for all-by-all BLAST (≥ 1; default: 5)

Neighborhood window size: 10 Number of neighbors to retrieve on either side of the query sequence for each BLAST result (default: 10)

E-mail address:

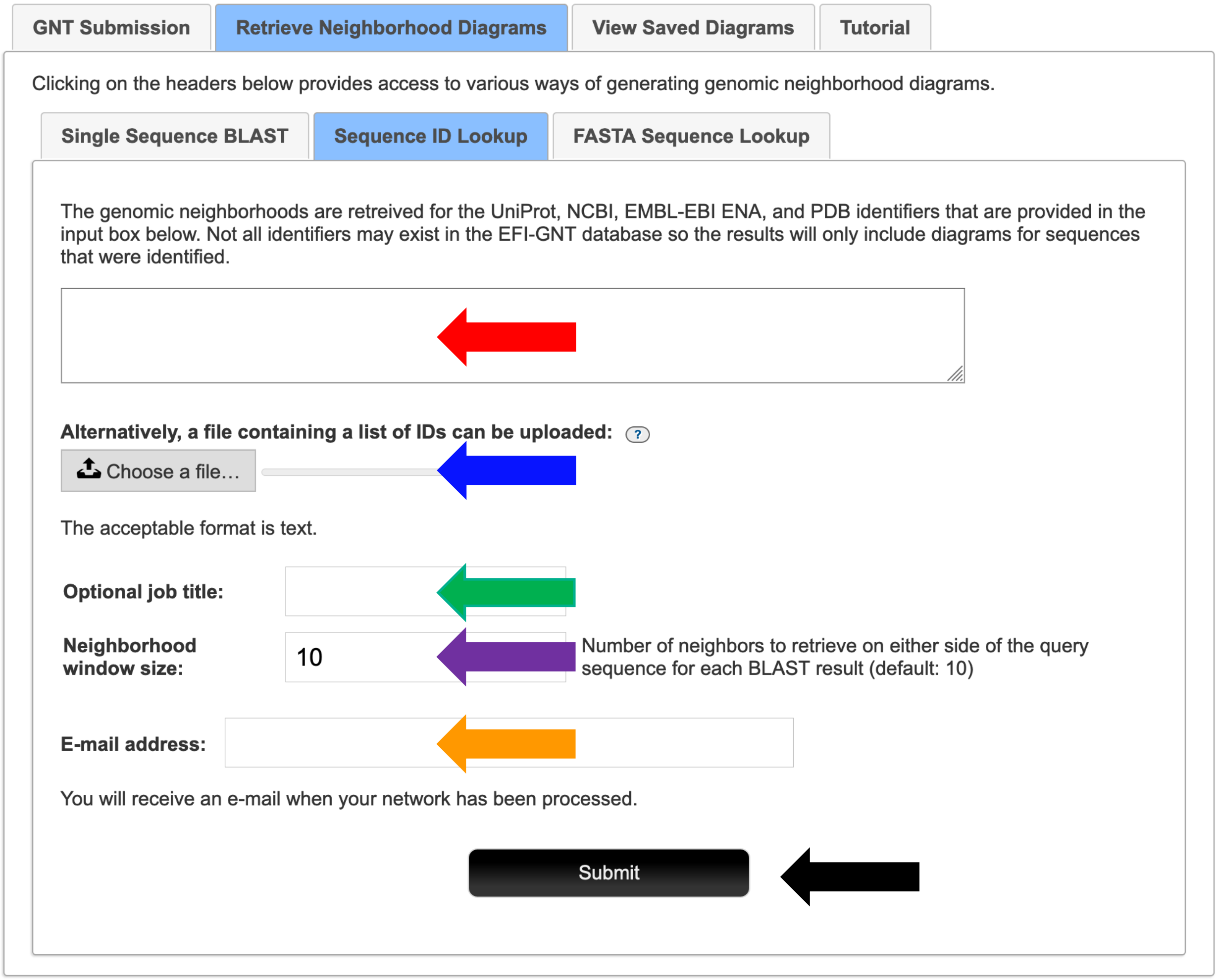
You will receive an e-mail when your network has been processed.

The number of sequences retrieved (default 200; ≤500; green arrow), negative log of the e-value for the BLAST (default 5, ≥1; magenta arrow), and neighborhood size (default ±10; from ±3 to ±20; orange arrow) can be specified. An "**Optional job title**" (blue arrow) can be provided that will appear on the "**Previous Jobs**" panel. On the "**Previous Jobs**" panel, these jobs are designated "**Sequence BLAST**".

An "**E-mail address**" (cyan arrow) is required. The job is initiated with the "**Submit**" button (black arrow).

The progress of the job is available on the "**Previous Jobs**" panel: "**PENDING**" when the job is received, "**RUNNING**" while the job is running, and a link to the results page when the job is finished. Also, an e-mail is sent to the user when the job is started; a second e-mail is sent when the job is finished that includes a link to the results page.

"Sequence ID Lookup" tab: The GNDs for one or more proteins specified by a list (entered in the box; red arrow) or uploaded as a file ("**Choose a file**"; blue arrow) of UniProt and/or NCBI IDs are retrieved; the UniProt-provided "idmapping" table is used to associate NCBI IDs to UniProt IDs (when possible; see Option C of EFI-EST).



Clicking on the headers below provides access to various ways of generating genomic neighborhood diagrams.

The genomic neighborhoods are retreived for the UniProt, NCBI, EMBL-EBI ENA, and PDB identifiers that are provided in the input box below. Not all identifiers may exist in the EFI-GNT database so the results will only include diagrams for sequences that were identified.

Alternatively, a file containing a list of IDs can be uploaded: ?

The acceptable format is text.

Optional job title:

Neighborhood window size: 10

Number of neighbors to retrieve on either side of the query sequence for each BLAST result (default: 10)

E-mail address:

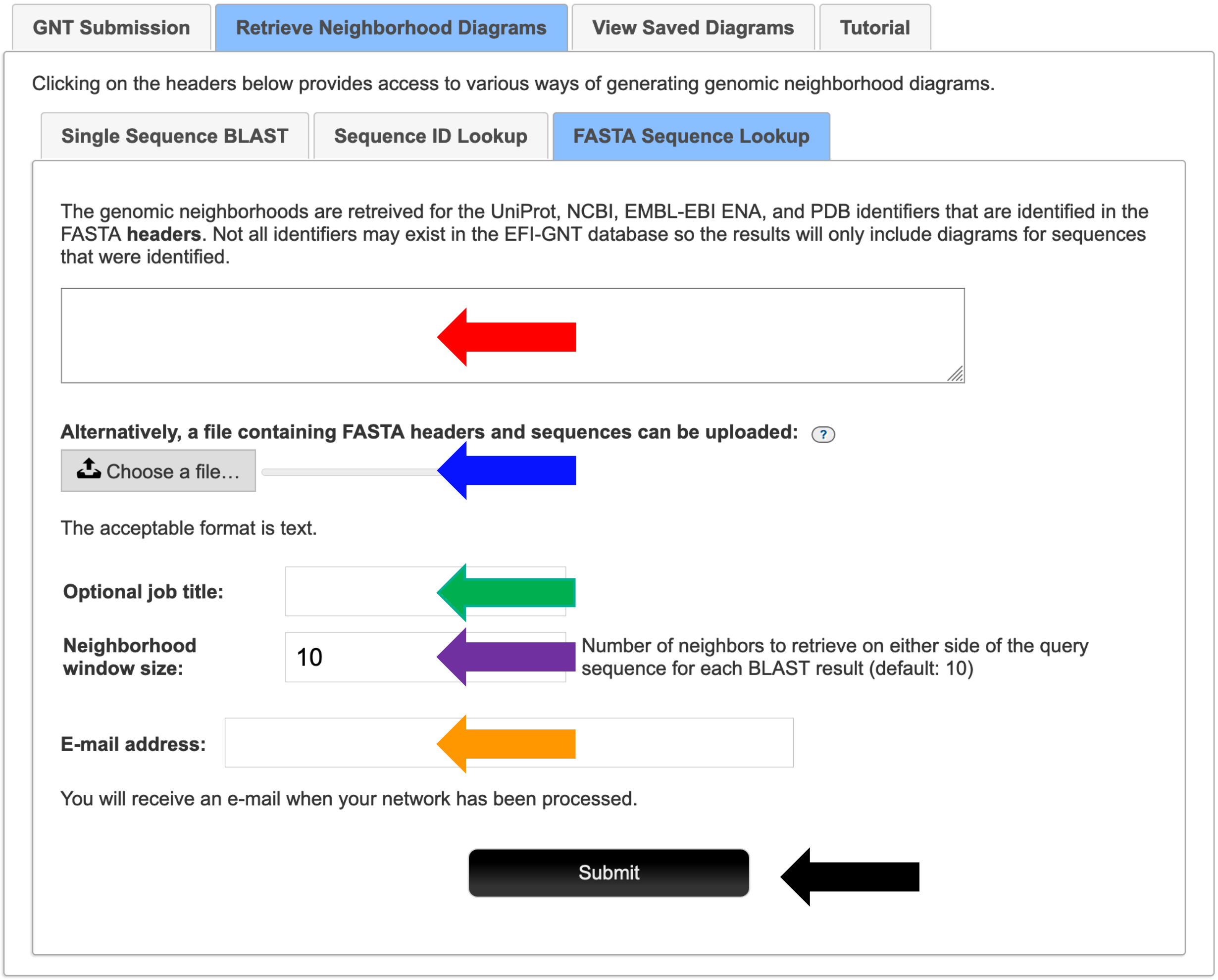
You will receive an e-mail when your network has been processed.

The genome neighborhood size (default ±10; from ±3 to ±20; magenta arrow) can be specified. An "**Optional job title**" (green arrow) can be provided that will appear on the "**Previous Jobs**" panel. On the "**Previous Jobs**" panel, these jobs are designated "**Sequence ID lookup**".

An “**E-mail address**” (orange arrow) is required. The job is initiated with the “**Submit**” button (black arrow).

The progress of the job is available on the “**Previous Jobs**” panel: “**PENDING**” when the job is received, “**RUNNING**” while the job is running, and a link to the results page when the job is finished. Also, an e-mail is sent to the user when the job is started; a second e-mail is sent when the job is finished that includes a link to the results page.

"FASTA Sequence Lookup" tab: The GNDs for one or more proteins specified in FASTA sequences (entered in the box; red arrow) or uploaded as a file ("**Choose a file**"; blue arrow); the FASTA headers are read for UniProt/NCBI accession IDs and the UniProt-provided "idmapping" table is used to associate NCBI IDs to UniProt IDs (when possible; see Option C of EFI-EST).



Clicking on the headers below provides access to various ways of generating genomic neighborhood diagrams.

The genomic neighborhoods are retreived for the UniProt, NCBI, EMBL-EBI ENA, and PDB identifiers that are identified in the FASTA **headers**. Not all identifiers may exist in the EFI-GNT database so the results will only include diagrams for sequences that were identified.

Alternatively, a file containing FASTA headers and sequences can be uploaded:

The acceptable format is text.

Optional job title:

Neighborhood window size: 10

Number of neighbors to retrieve on either side of the query sequence for each BLAST result (default: 10)

E-mail address:

You will receive an e-mail when your network has been processed.

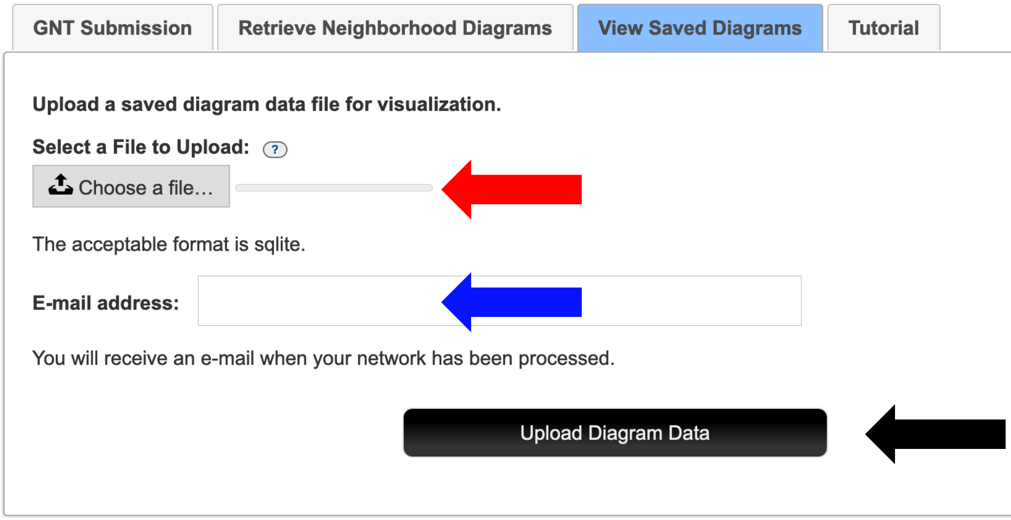
The genome neighborhood size (default ±10; from ±3 to ±20; magenta arrow) can be specified. An "**Optional job title**" (green arrow) can be provided that will appear on the "**Previous Jobs**" panel. On the "**Previous Jobs**" panel, these jobs are designated "**FASTA header ID lookup**".

An “**E-mail address**” (orange arrow) is required. The job is initiated with the “**Submit**” button (black arrow).

The progress of the job is available on the “**Previous Jobs**” panel: “**PENDING**” when the job is received, “**RUNNING**” while the job is running, and a link to the results page when the job is finished. Also, an e-mail is sent to the user when the job is started; a second e-mail is sent when the job is finished that includes a link to the results page.

D. EFI-GNT Home Page, “View Saved Diagrams” tab: Upload Saved GNDs from a previous session

GNDs that were downloaded from a previous job can be uploaded (“**Choose a file**”; red arrow) and viewed.



Upload a saved diagram data file for visualization.

Select a File to Upload:

The acceptable format is sqlite.

E-mail address:

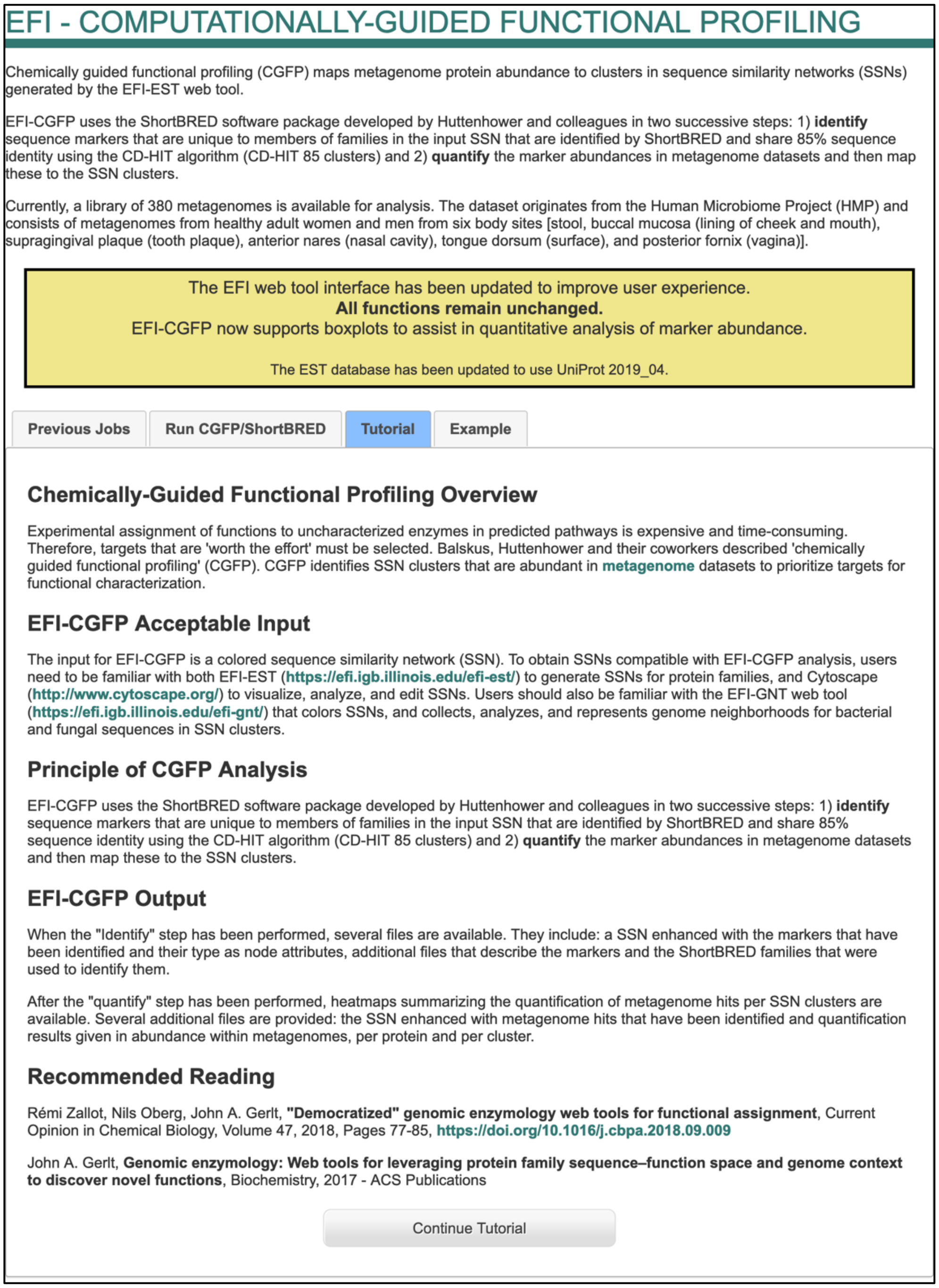
You will receive an e-mail when your network has been processed.

An “**E-mail address**” (blue arrow) is required. The job is initiated with the “**Upload Diagram Data**” button (black arrow). On the “**Previous Jobs**” panel, these jobs are designated “**Upload diagram data file**”.

The progress of the job is available on the “**Previous Jobs**” panel: “**PENDING**” when the job is received, “**RUNNING**” while the job is running, and a link to the results page when the job is finished. Also, an e-mail is sent to the user when the job is started; a second e-mail is sent when the job is finished that includes a link to the results page.

IV. EFI-CGFP (https://efi.igb.illinois.edu/efi-cgfp/)

This section describes EFI-CGFP, including the options available for identifying markers for the clusters in the input SSNs, the metagenomes that can be selected for marker/cluster abundance quantification, and the heatmaps/boxplots that are available for download.



EFI - COMPUTATIONALLY-GUIDED FUNCTIONAL PROFILING

Chemically guided functional profiling (CGFP) maps metagenome protein abundance to clusters in sequence similarity networks (SSNs) generated by the EFI-EST web tool.

EFI-CGFP uses the ShortBRED software package developed by Huttenhower and colleagues in two successive steps: 1) **identify** sequence markers that are unique to members of families in the input SSN that are identified by ShortBRED and share 85% sequence identity using the CD-HIT algorithm (CD-HIT 85 clusters) and 2) **quantify** the marker abundances in metagenome datasets and then map these to the SSN clusters.

Currently, a library of 380 metagenomes is available for analysis. The dataset originates from the Human Microbiome Project (HMP) and consists of metagenomes from healthy adult women and men from six body sites [stool, buccal mucosa (lining of cheek and mouth), supragingival plaque (tooth plaque), anterior nares (nasal cavity), tongue dorsum (surface), and posterior fornix (vagina)].

The EFI web tool interface has been updated to improve user experience.
All functions remain unchanged.
EFI-CGFP now supports boxplots to assist in quantitative analysis of marker abundance.

The EST database has been updated to use UniProt 2019_04.

Chemically-Guided Functional Profiling Overview

Experimental assignment of functions to uncharacterized enzymes in predicted pathways is expensive and time-consuming. Therefore, targets that are 'worth the effort' must be selected. Balskus, Huttenhower and their coworkers described 'chemically guided functional profiling' (CGFP). CGFP identifies SSN clusters that are abundant in metagenome datasets to prioritize targets for functional characterization.

EFI-CGFP Acceptable Input

The input for EFI-CGFP is a colored sequence similarity network (SSN). To obtain SSNs compatible with EFI-CGFP analysis, users need to be familiar with both EFI-EST (**https://efi.igb.illinois.edu/efi-est/**) to generate SSNs for protein families, and Cytoscape (**http://www.cytoscape.org/**) to visualize, analyze, and edit SSNs. Users should also be familiar with the EFI-GNT web tool (**https://efi.igb.illinois.edu/efi-gnt/**) that colors SSNs, and collects, analyzes, and represents genome neighborhoods for bacterial and fungal sequences in SSN clusters.

Principle of CGFP Analysis

EFI-CGFP uses the ShortBRED software package developed by Huttenhower and colleagues in two successive steps: 1) **identify** sequence markers that are unique to members of families in the input SSN that are identified by ShortBRED and share 85% sequence identity using the CD-HIT algorithm (CD-HIT 85 clusters) and 2) **quantify** the marker abundances in metagenome datasets and then map these to the SSN clusters.

EFI-CGFP Output

When the "Identify" step has been performed, several files are available. They include: a SSN enhanced with the markers that have been identified and their type as node attributes, additional files that describe the markers and the ShortBRED families that were used to identify them.

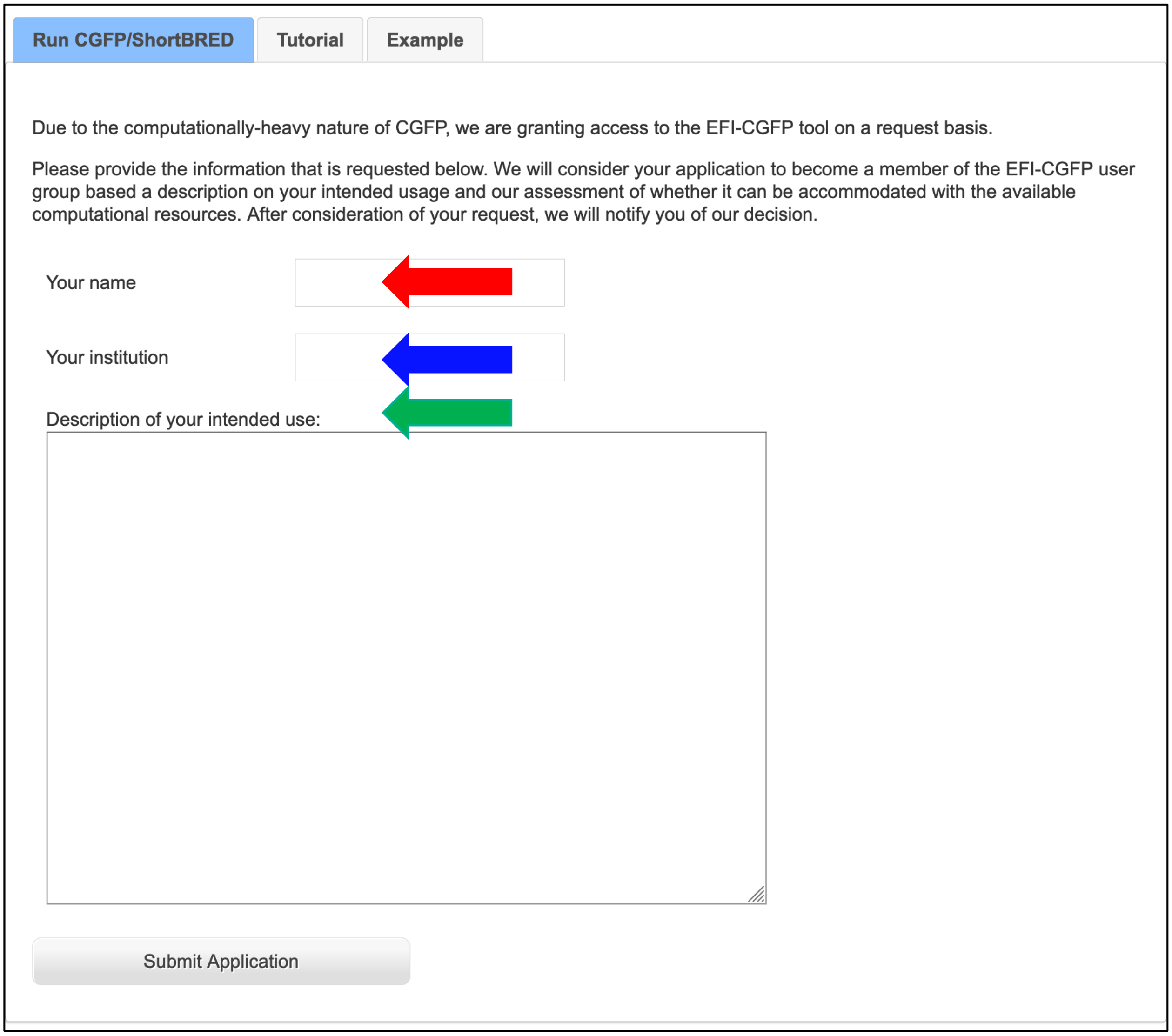
After the "quantify" step has been performed, heatmaps summarizing the quantification of metagenome hits per SSN clusters are available. Several additional files are provided: the SSN enhanced with metagenome hits that have been identified and quantification results given in abundance within metagenomes, per protein and per cluster.

Recommended Reading

Rémi Zallot, Nils Oberg, John A. Gerlt, **"Democratized" genomic enzymology web tools for functional assignment**, Current Opinion in Chemical Biology, Volume 47, 2018, Pages 77-85, **https://doi.org/10.1016/j.cbpa.2018.09.009**

John A. Gerlt, **Genomic enzymology: Web tools for leveraging protein family sequence–function space and genome context to discover novel functions**, Biochemistry, 2017 - ACS Publications

EFI-CGFP User Group. Given the computational resources required for both marker identification and metagenome quantification, use of EFI-CGFP requires that the user be an approved member of the EFI-CGFP user group. The first time the "**Run CGFP/ShortBRED**" page is accessed by a registered user, the user is asked to apply for membership in the EFI-CGFP user group. The user provides his/her name (red arrow), institution (blue arrow), and a brief description of the proposed CGFP project (green arrow), most critically the size of the protein family because this impacts the computational requirements. We will provide feedback to the user, including advice in using EFI-CGFP.



Due to the computationally-heavy nature of CGFP, we are granting access to the EFI-CGFP tool on a request basis.

Please provide the information that is requested below. We will consider your application to become a member of the EFI-CGFP user group based a description on your intended usage and our assessment of whether it can be accommodated with the available computational resources. After consideration of your request, we will notify you of our decision.

Your name

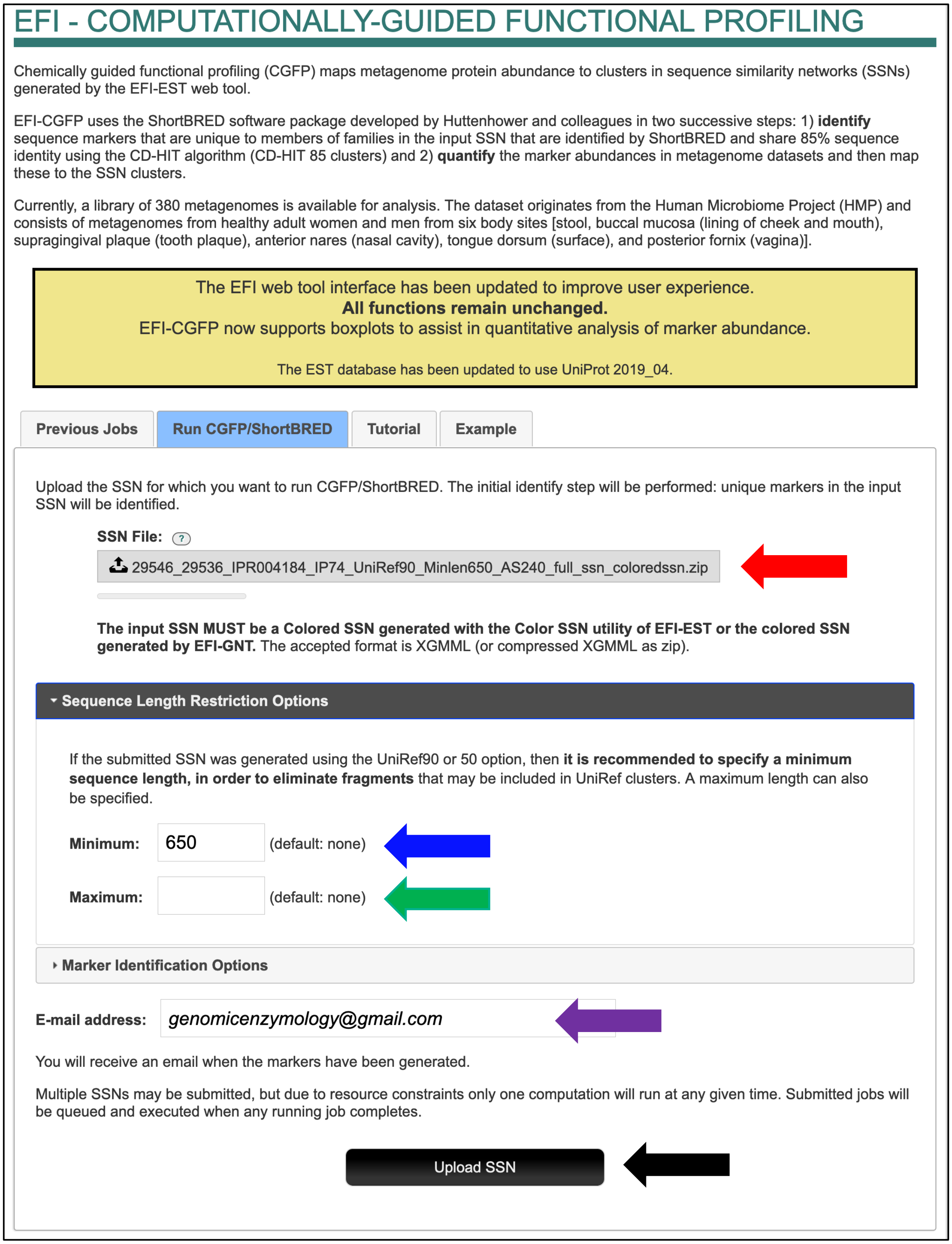
Your institution

Description of your intended use:

Overview of the CGFP pipeline. We developed EFI-CGFP to make CGFP[9] "user friendly", i.e., easily accessible to experimentalists. The user uploads a colored SSN (from either the Color SSNs utility or EFI-GNT) with numbered clusters and singletons that are required for marker identification. When marker identification is complete using the ShortBRED algorithm[10], the user chooses metagenome datasets for abundance determination. When marker abundance determination is complete, the user can access heat maps and box plots that displays metagenome abundance for each SSN cluster and singleton; an SSN is available that includes the results of marker identification and metagenome abundance.

Metagenome library for CGFP. The metagenome library includes 380 datasets from healthy adult men and women (Human Microbiome Project; HMP). The metagenomes were collected from six body sites: anterior nares (56 datasets), buccal mucosa (66 datasets), posterior fornix (31 datasets), stool (82 datasets), supragingival plaque (72 datasets), and tongue dorsum (73 datasets). We anticipate that additional libraries, e.g., environmental metagenome datasets, will be added in response to the user's requests.

A. EFI-CGFP Home Page, Marker Identification ("Run CGFP/ShortBRED" tab)



EFI - COMPUTATIONALLY-GUIDED FUNCTIONAL PROFILING

Chemically guided functional profiling (CGFP) maps metagenome protein abundance to clusters in sequence similarity networks (SSNs) generated by the EFI-EST web tool.

EFI-CGFP uses the ShortBRED software package developed by Huttenhower and colleagues in two successive steps: 1) **identify** sequence markers that are unique to members of families in the input SSN that are identified by ShortBRED and share 85% sequence identity using the CD-HIT algorithm (CD-HIT 85 clusters) and 2) **quantify** the marker abundances in metagenome datasets and then map these to the SSN clusters.

Currently, a library of 380 metagenomes is available for analysis. The dataset originates from the Human Microbiome Project (HMP) and consists of metagenomes from healthy adult women and men from six body sites [stool, buccal mucosa (lining of cheek and mouth), supragingival plaque (tooth plaque), anterior nares (nasal cavity), tongue dorsum (surface), and posterior fornix (vagina)].

The EFI web tool interface has been updated to improve user experience.
All functions remain unchanged.
EFI-CGFP now supports boxplots to assist in quantitative analysis of marker abundance.

The EST database has been updated to use UniProt 2019_04.

Upload the SSN for which you want to run CGFP/ShortBRED. The initial identify step will be performed: unique markers in the input SSN will be identified.

SSN File: ?
29546_29536_IPR004184_IP74_UniRef90_Minlen650_AS240_full_ssn_coloredssn.zip

The input SSN MUST be a Colored SSN generated with the Color SSN utility of EFI-EST or the colored SSN generated by EFI-GNT. The accepted format is XGMML (or compressed XGMML as zip).

Sequence Length Restriction Options

If the submitted SSN was generated using the UniRef90 or 50 option, then **it is recommended to specify a minimum sequence length, in order to eliminate fragments** that may be included in UniRef clusters. A maximum length can also be specified.

Minimum: 650 (default: none)

Maximum: (default: none)

Marker Identification Options

E-mail address: *genomicenzymology@gmail.com*

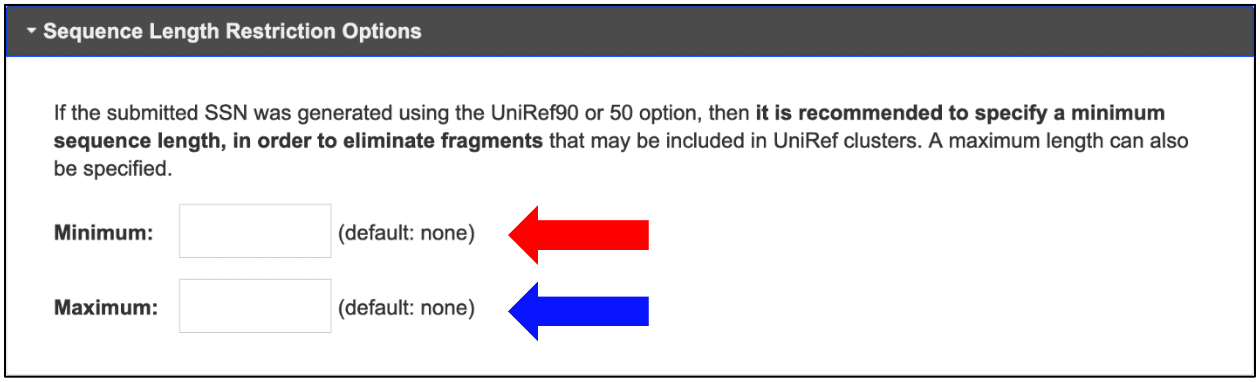
You will receive an email when the markers have been generated.

Multiple SSNs may be submitted, but due to resource constraints only one computation will run at any given time. Submitted jobs will be queued and executed when any running job completes.

The xgmml file for a colored SSN is uploaded ("**SSN File**"; generated with either the Color SSNs utility of EFI-EST or EFI-GNT; uncompressed or zipped; red arrow); the SSN can be generated for either full length sequences or domains.

The default parameters for marker identification can be changed using the options in the accordion windows below the upload for the input sequence: "**Sequence Length Restriction Options**" and "**Marker Identification Options**".

Sequence Length Restriction Options. If the desired length restrictions were not used to generate the SSN, we recommend entering a value in the "**Minimum**" box (red arrow) for the minimum sequence length that corresponds to the lower limit for full length sequences in the input SSN to ensure that the consensus sequences for the ShortBRED families are not biased by fragments (the length filter is applied to the full set of sequences, including members of UniRef clusters, if the SSN was generated from the UniRef database).



▾ **Sequence Length Restriction Options**

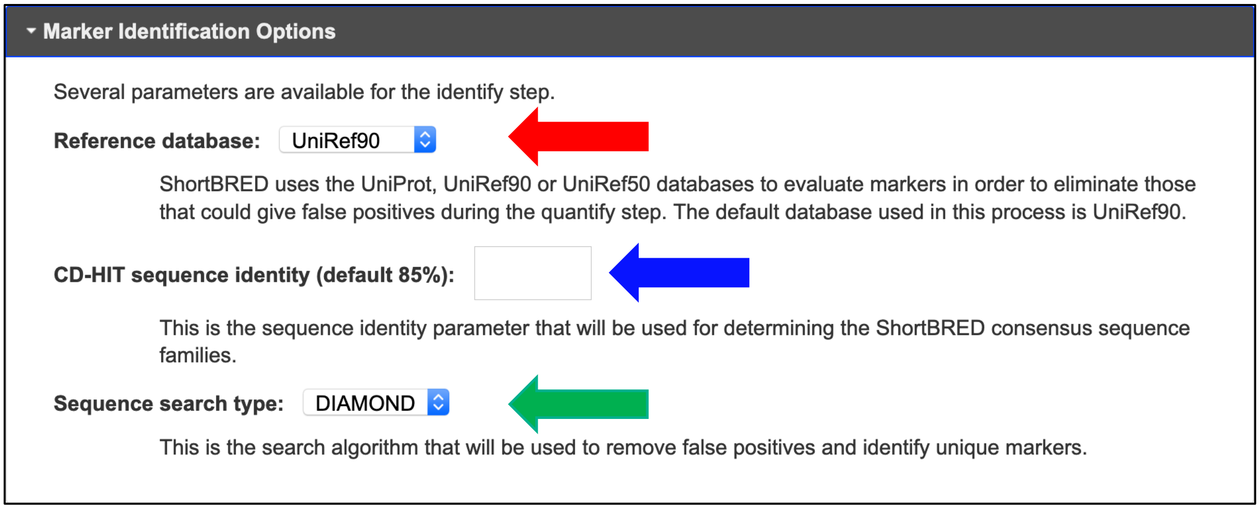
If the submitted SSN was generated using the UniRef90 or 50 option, then **it is recommended to specify a minimum sequence length, in order to eliminate fragments** that may be included in UniRef clusters. A maximum length can also be specified.

Minimum: (default: none)

Maximum: (default: none)

The nodes in SSNs generated using UniRef50 and UniRef90 clusters will contain sequences that are ≥80% the length of the seed sequence; therefore, some of these clusters will contain "truncated" sequences. When using UniRef SSNs, the user should use length histogram for UniProt IDs provided on the "Dataset Completed" page for the initial SSN job to choose the value for the minimum sequence length for full-length proteins. If sequence length restrictions were used to generate the SSN, they need not be reapplied.

Marker Identification Options. EFI-CGFP uses ShortBRED for family marker identification; it defines "ShortBRED families" that are distinct from SSN families (e.g., isofunctional clusters). CD-HIT[11, 12] clusters the unique sequences in the input SSN (using all UniProt IDs in the metanodes in UniRef clusters and rep node SSNs after the recommended length filtering); unique sequences are used so that the consensus sequences used for marker identification are not biased by multiple occurrences of the same sequence. The default sequence identity for CD-HIT is 85%; the user can change this parameter by entering an alternate value in the "**CD-HIT sequence identity**" box (blue arrow). For example, if the user is aware that function is critically dependent on sequence identity in the input SSN, a larger percent identity should be used to ensure that the ShortBRED families are isofunctional. The sequences within each ShortBRED family then are aligned and a consensus sequence is determined using MUSCLE.



Marker Identification Options

Several parameters are available for the identify step.

Reference database: UniRef90

ShortBRED uses the UniProt, UniRef90 or UniRef50 databases to evaluate markers in order to eliminate those that could give false positives during the quantify step. The default database used in this process is UniRef90.

CD-HIT sequence identity (default 85%):

This is the sequence identity parameter that will be used for determining the ShortBRED consensus sequence families.

Sequence search type: DIAMOND

This is the search algorithm that will be used to remove false positives and identify unique markers.

Marker identification involves pairwise comparisons of 1) the consensus sequences for all ShortBRED families among themselves and 2) then the consensus sequences against a reference database to both identify ShortBRED family-specific markers and, also, remove false positives. The default reference database is the complete set of UniRef90 cluster IDs (ShortBRED markers should not identify nonhomologous families); the complete UniProt database or the complete set of UniRef50 cluster IDs can be selected with the "**Reference database**" pull-down menu (red arrow). The execution time with the UniProt database is approximately twice that for UniRef90 but excludes more false positives; the execution time for the UniRef50 is approximately one-half that for UniRef90 but includes more false positives. As a compromise, we recommend using the UniRef90 default, especially for jobs with a large number of UniProt IDs.

Finally, DIAMOND[13] is the default comparison algorithm used by EFI-CGFP, although the user can select BLAST in the "**Sequence search type**" pull-down menu (green arrow). BLAST is more accurate but considerably slower than DIAMOND (≥30-fold difference in execution time). For "small" families, e.g., the glycyl radical enzyme superfamily (IPR004184; ~11,000 sequences) described later in this article, marker identification using BLAST requires 24 hrs but marker identification using DIAMOND requires 45 min; for "large" families, e.g., the radical SAM superfamily (IPR007197 and IPR006638; ~475,000 sequences), marker identification using BLAST requires one month but marker identification using DIAMOND requires 10 hrs. A compromise of speed and accuracy is required: we recommend the initial use of DIAMOND for a family SSN and then, if desired, a subsequent analysis using BLAST of clusters of interest in a daughter SSN generated using Cytoscape.

On the “**Previous Jobs**” panel and the “**Job History**” page, these jobs are designated “**CGFP Identification**”.

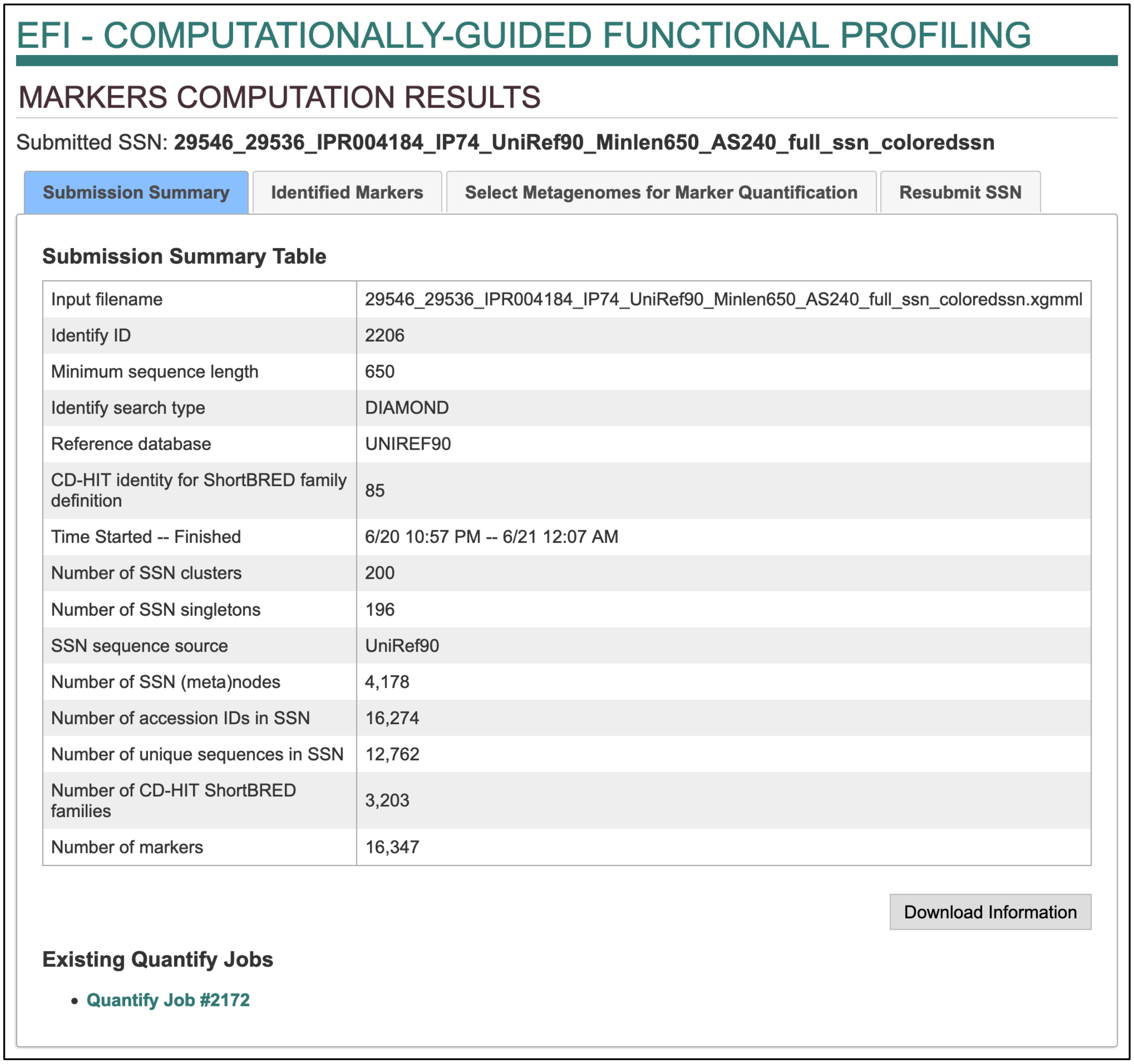
An “**E-mail address**” (magenta arrow) is required. The job is initiated with the “**Submit Analysis**” button (black arrow).

The progress of the job is available on the “**Previous Jobs**” panel: “**PENDING**” when the job is received, “**RUNNING**” while the job is running, and a link to the results page when the job is finished. Also, an e-mail is sent to the user when the job is started; a second e-mail is sent when the job is finished that includes a link to the results page.

Because marker identification jobs can be time-consuming, the user can stop a job by clicking the red button adjacent to its RUNNING status notation in the “**Previous Jobs**” panel.

B. EFI-CGFP "Markers Computation Results" page

This page is available when marker identification is completed.



EFI - COMPUTATIONALLY-GUIDED FUNCTIONAL PROFILING
MARKERS COMPUTATION RESULTS
Submission Summary
Identified Markers
Select Metagenomes for Marker Quantification
Resubmit SSN
Download Information

Submitted SSN: **29546_29536_IPR004184_IP74_UniRef90_Minlen650_AS240_full_ssn_coloredssn**

Submission Summary Table

Input filename	29546_29536_IPR004184_IP74_UniRef90_Minlen650_AS240_full_ssn_coloredssn.xgmml
Identify ID	2206
Minimum sequence length	650
Identify search type	DIAMOND
Reference database	UNIREF90
CD-HIT identity for ShortBRED family definition	85
Time Started -- Finished	6/20 10:57 PM -- 6/21 12:07 AM
Number of SSN clusters	200
Number of SSN singletons	196
SSN sequence source	UniRef90
Number of SSN (meta)nodes	4,178
Number of accession IDs in SSN	16,274
Number of unique sequences in SSN	12,762
Number of CD-HIT ShortBRED families	3,203
Number of markers	16,347

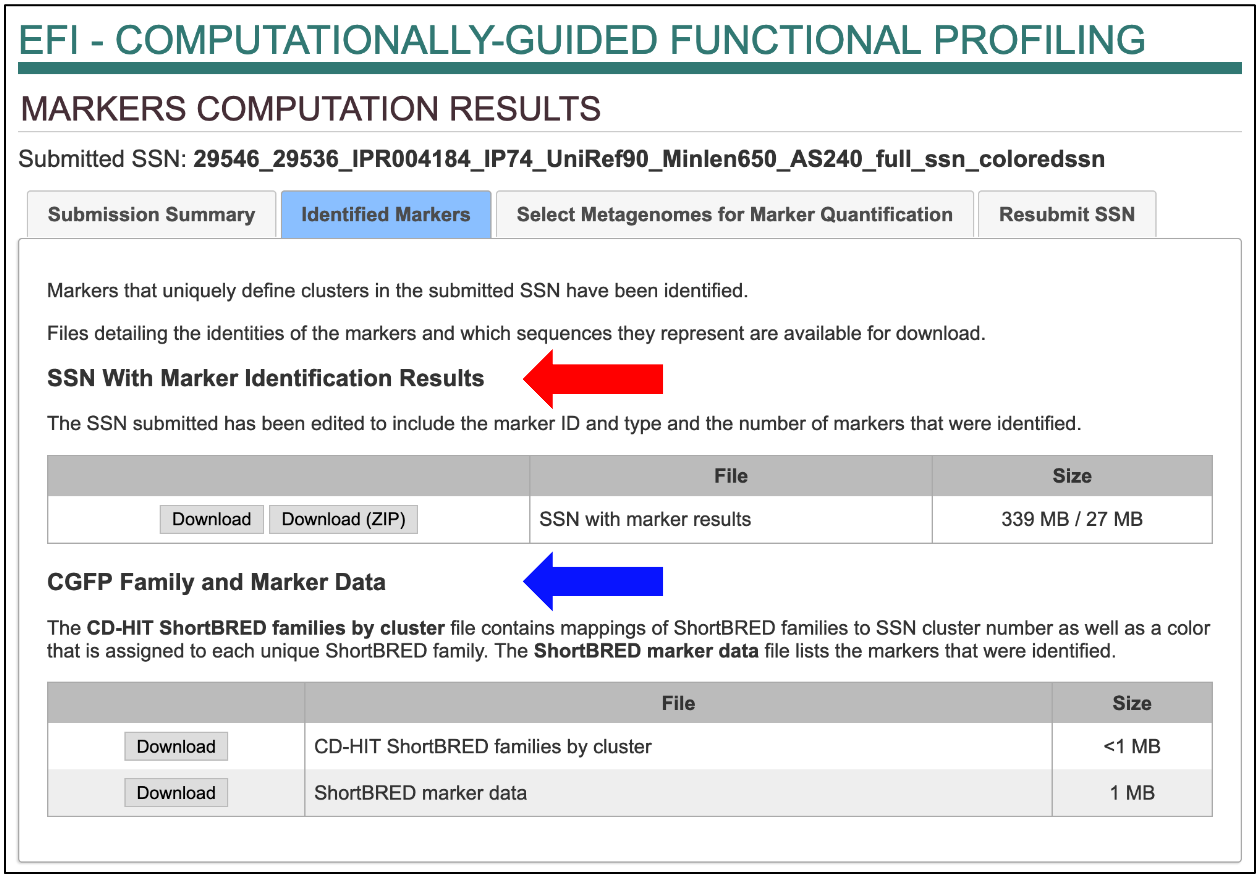
Existing Quantify Jobs

- Quantify Job #2172

"Submission Summary" tab. This tab provides a table summarizing the job parameters and information about the SSN file, including the number of SSN clusters, singletons, (meta)nodes, accession IDs, unique sequences, ShortBRED families that were identified, and family markers that can be downloaded as a text file for future reference.

This page also includes a list (with links) to marker quantification/metagenome abundance jobs that were performed with the computed markers.

"Identified Markers" tab. The "**SSN with Marker Identification Results**" panel (red arrow) tab provides access to the input SSN to which four node attributes are added ("SSN with marker results"): "Seed Sequences" [the (meta)node contains the UniProt ID for the seed sequence for a ShortBRED family], "Seed Sequence Cluster(s)" [the Seed Sequence to which the (meta)node contributes family members] "Marker Types" (true, quasi, or junction), and "Number of Markers". The "**CGFP Family and Marker Data**" panel (blue arrow) provides access to text files that define the membership of the ShortBRED families ("**CD-HIT ShortBRED families by clusters**") and provide the sequences of the markers determined for the ShortBRED family consensus sequences. ("**ShortBRED marker data**"). The additional node attributes in the SSNs generated by the marker computation step of EFI-CGFP are listed in Table S6.



EFI - COMPUTATIONALLY-GUIDED FUNCTIONAL PROFILING

MARKERS COMPUTATION RESULTS

Submitted SSN: **29546_29536_IPR004184_IP74_UniRef90_Minlen650_AS240_full_ssn_coloredssn**

Markers that uniquely define clusters in the submitted SSN have been identified.

Files detailing the identities of the markers and which sequences they represent are available for download.

SSN With Marker Identification Results

The SSN submitted has been edited to include the marker ID and type and the number of markers that were identified.

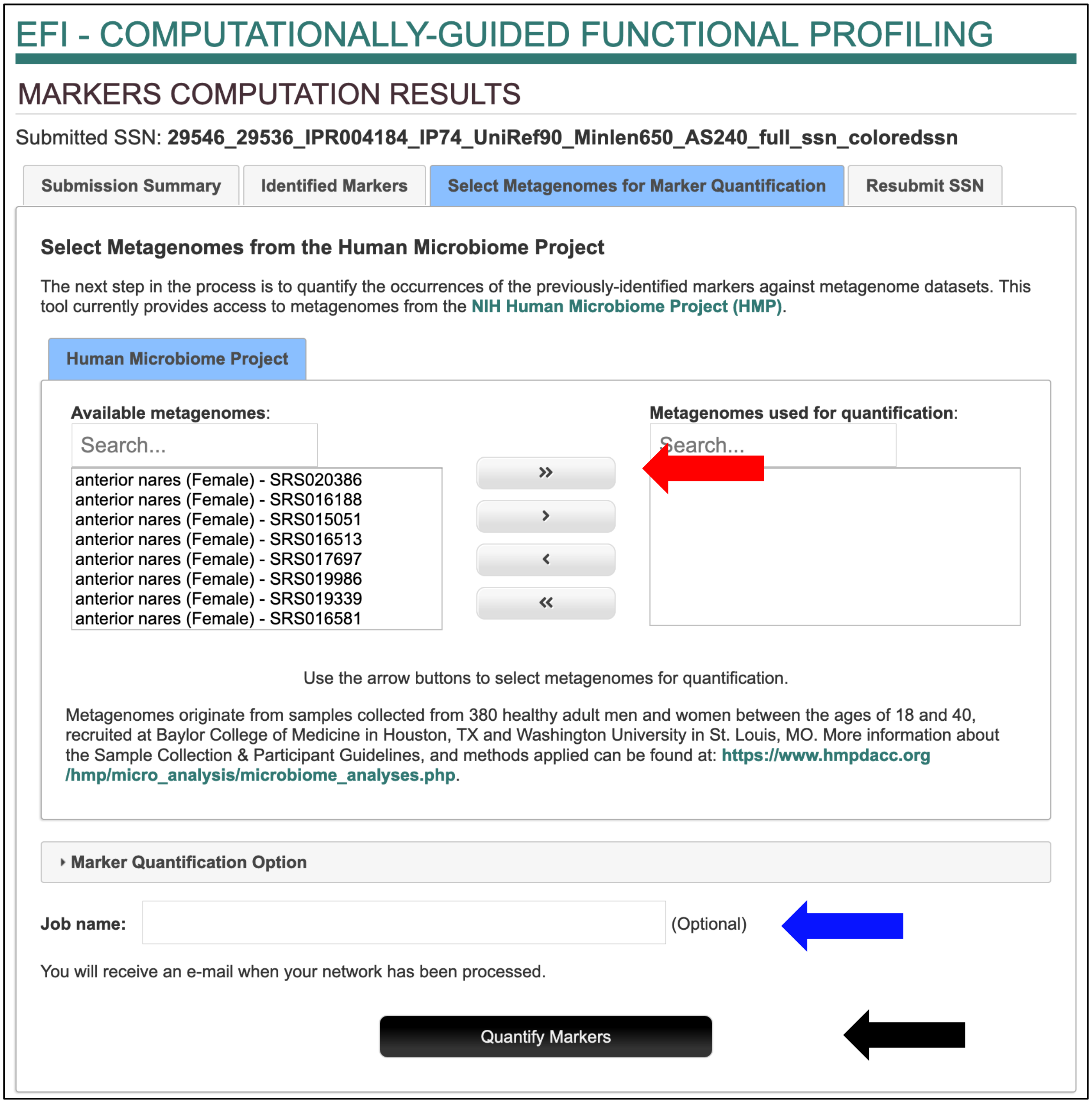
	File	Size
Download Download (ZIP)	SSN with marker results	339 MB / 27 MB

CGFP Family and Marker Data

The **CD-HIT ShortBRED families by cluster** file contains mappings of ShortBRED families to SSN cluster number as well as a color that is assigned to each unique ShortBRED family. The **ShortBRED marker data** file lists the markers that were identified.

	File	Size
Download	CD-HIT ShortBRED families by cluster	<1 MB
Download	ShortBRED marker data	1 MB

"Select Metagenomes for Marker Quantification" tab. This tab provides a library of 380 healthy metagenome datasets from six body sites ("**Human Microbiome Project**"); the user can select (red arrow) all members or a subset of the library for metagenome abundance quantification (by entering text in the "**Search**" window, e.g., "stool", "anterior nares", or "male"). We expect that the types and number of metagenome databases will increase with demand.



EFI - COMPUTATIONALLY-GUIDED FUNCTIONAL PROFILING

MARKERS COMPUTATION RESULTS

Submitted SSN: **29546_29536_IPR004184_IP74_UniRef90_Minlen650_AS240_full_ssn_coloredssn**

Select Metagenomes from the Human Microbiome Project

The next step in the process is to quantify the occurrences of the previously-identified markers against metagenome datasets. This tool currently provides access to metagenomes from the **NIH Human Microbiome Project (HMP)**.

Available metagenomes:

anterior nares (Female) - SRS020386
anterior nares (Female) - SRS016188
anterior nares (Female) - SRS015051
anterior nares (Female) - SRS016513
anterior nares (Female) - SRS017697
anterior nares (Female) - SRS019986
anterior nares (Female) - SRS019339
anterior nares (Female) - SRS016581

Metagenomes used for quantification:

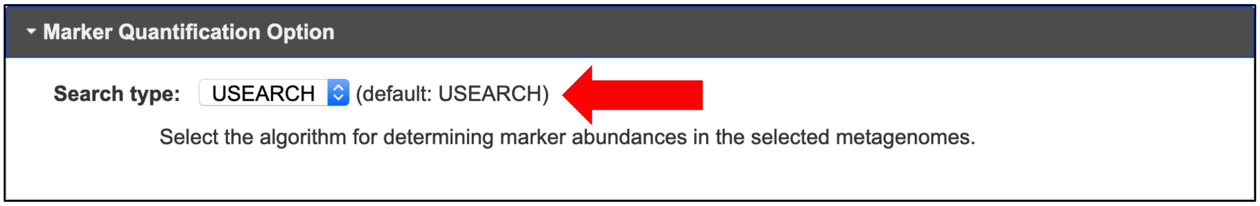
Use the arrow buttons to select metagenomes for quantification.

Metagenomes originate from samples collected from 380 healthy adult men and women between the ages of 18 and 40, recruited at Baylor College of Medicine in Houston, TX and Washington University in St. Louis, MO. More information about the Sample Collection & Participant Guidelines, and methods applied can be found at: **https://www.hmpdacc.org /hmp/micro_analysis/microbiome_analyses.php**.

Job name: (Optional)

You will receive an e-mail when your network has been processed.

With the "**Marker Quantification Option**" pull-down menu (red arrow) the user can select either USEARCH[14] (default) or DIAMOND (user-specified option) for the pairwise comparisons of the markers with the metagenome datasets for determination of metagenome abundance. Although DIAMOND is marginally faster than USEARCH, we recommend the default use of USEARCH because it is more accurate.



▾ Marker Quantification Option

Search type: USEARCH (default: USEARCH)

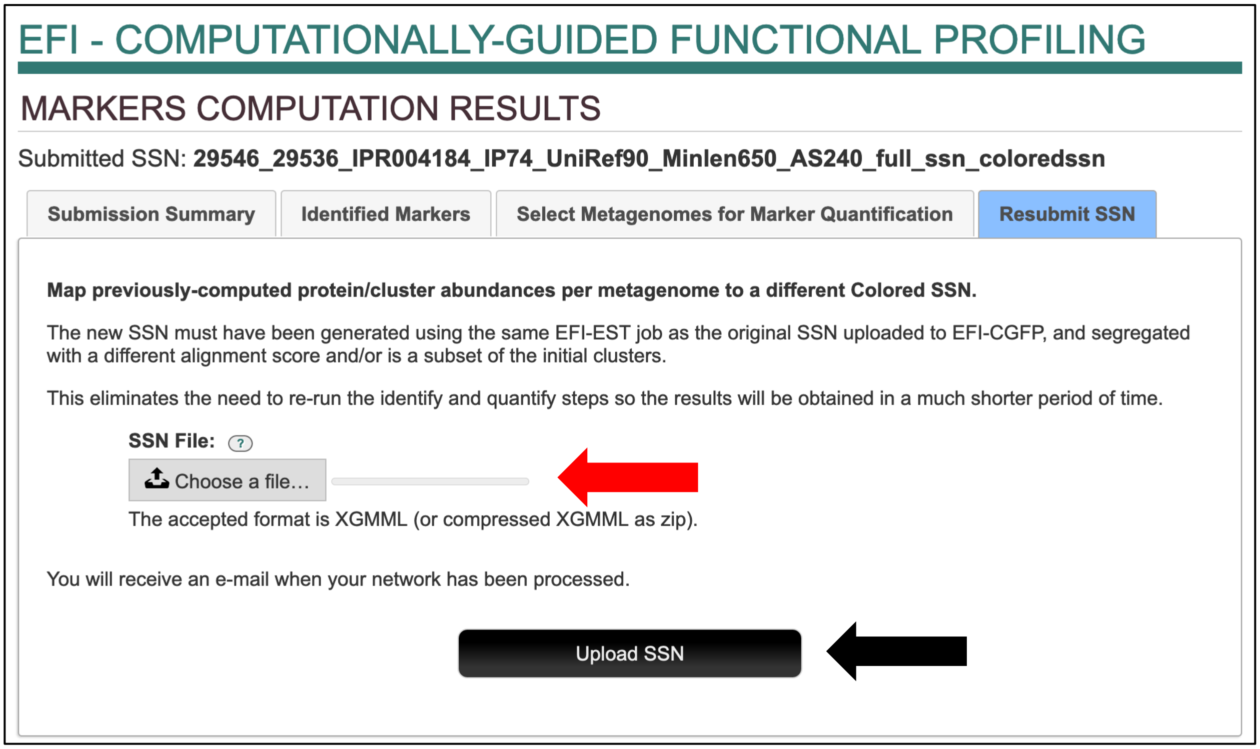
Select the algorithm for determining marker abundances in the selected metagenomes.

An optional "**Job name**" can be entered (blue arrow) that will appear on the "**Previous Jobs**" panel and on the "**Job History**" page. On the "**Previous Jobs**" panel and the "**Job History**" page, these jobs are designated "**CGFP Identification**".

The job is initiated with the "**Quantify Markers** button (black arrow).

The progress of the job is available on the "**Previous Jobs**" panel: "**PENDING**" when the job is received, "**RUNNING**" while the job is running, and a link to the results page when the job is finished. Also, an e-mail is sent to the user when the job is started; a second e-mail is sent when the job is finished that includes a link to the results page.

"Resubmit SSN" tab. After marker identification and metagenome quantification have been completed, this tab will be available so that the user can upload ("**SSN File, Choose a file:**"; red arrow) a colored SSN with the same sequences but generated with a different alignment score or a daughter SSN of the input colored SSN that contains a subset of the sequences/clusters.


EFI - COMPUTATIONALLY-GUIDED FUNCTIONAL PROFILING
MARKERS COMPUTATION RESULTS
Submission Summary
Identified Markers
Select Metagenomes for Marker Quantification
Resubmit SSN
Choose a file…
Upload SSN

Submitted SSN: **29546_29536_IPR004184_IP74_UniRef90_Minlen650_AS240_full_ssn_coloredssn**

Map previously-computed protein/cluster abundances per metagenome to a different Colored SSN.

The new SSN must have been generated using the same EFI-EST job as the original SSN uploaded to EFI-CGFP, and segregated with a different alignment score and/or is a subset of the initial clusters.

This eliminates the need to re-run the identify and quantify steps so the results will be obtained in a much shorter period of time.

SSN File:

The accepted format is XGMML (or compressed XGMML as zip).

You will receive an e-mail when your network has been processed.

Because the metagenome abundances are determined for ShortBRED family consensus sequence markers, these are independent of the clustering in the input SSN, so the values will be the same as those determined in the initial analysis; therefore, generation of the child heatmaps and SSNs is much faster than the initial analysis. Before uploading, these SSNs must be processed with the Color SSNs utility to assign cluster and singleton numbers.

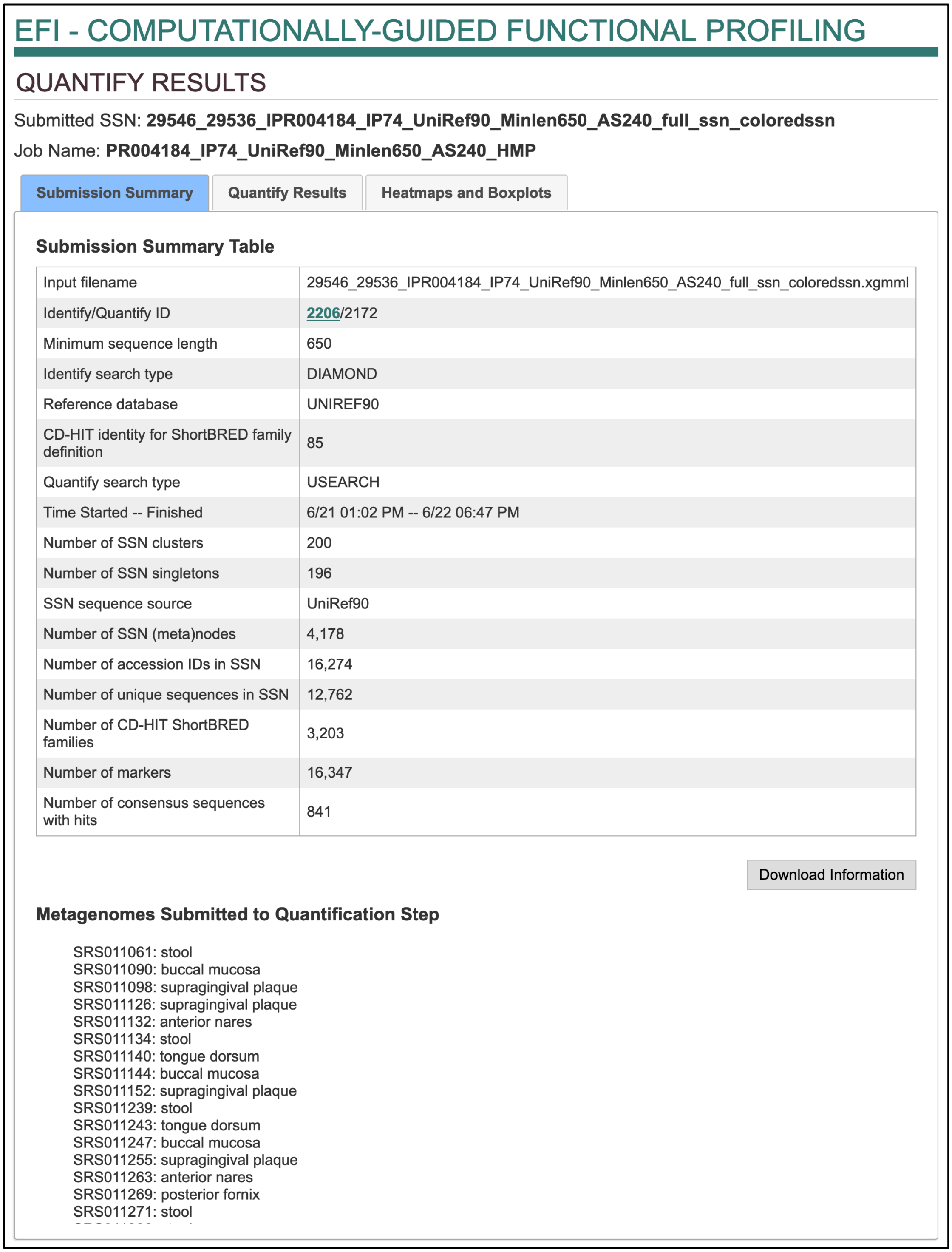
On the "**Previous Jobs**" panel and the "**Job History**" page, these jobs are designated "**CGFP Quantification**".

The job is initiated with the “**Upload SSN**” button (black arrow).

The progress of the job is available on the “**Previous Jobs**” panel: “**PENDING**” when the job is received, “**RUNNING**” while the job is running, and a link to the results page when the job is finished. Also, an e-mail is sent to the user when the job is started; a second e-mail is sent when the job is finished that includes a link to the results page.

Because metagenome abundance jobs can be time-consuming, the user can stop a job by clicking the red button adjacent to its RUNNING status notation.

C. EFI-CGFP "Quantify Results" page



EFI - COMPUTATIONALLY-GUIDED FUNCTIONAL PROFILING

QUANTIFY RESULTS

Submitted SSN: **29546_29536_IPR004184_IP74_UniRef90_Minlen650_AS240_full_ssn_coloredssn**

Job Name: **PR004184_IP74_UniRef90_Minlen650_AS240_HMP**

Submission Summary Table

Input filename	29546_29536_IPR004184_IP74_UniRef90_Minlen650_AS240_full_ssn_coloredssn.xgmml
Identify/Quantify ID	2206/2172
Minimum sequence length	650
Identify search type	DIAMOND
Reference database	UNIREF90
CD-HIT identity for ShortBRED family definition	85
Quantify search type	USEARCH
Time Started -- Finished	6/21 01:02 PM -- 6/22 06:47 PM
Number of SSN clusters	200
Number of SSN singletons	196
SSN sequence source	UniRef90
Number of SSN (meta)nodes	4,178
Number of accession IDs in SSN	16,274
Number of unique sequences in SSN	12,762
Number of CD-HIT ShortBRED families	3,203
Number of markers	16,347
Number of consensus sequences with hits	841

Metagenomes Submitted to Quantification Step

SRS011061: stool
SRS011090: buccal mucosa
SRS011098: supragingival plaque
SRS011126: supragingival plaque
SRS011132: anterior nares
SRS011134: stool
SRS011140: tongue dorsum
SRS011144: buccal mucosa
SRS011152: supragingival plaque
SRS011239: stool
SRS011243: tongue dorsum
SRS011247: buccal mucosa
SRS011255: supragingival plaque
SRS011263: anterior nares
SRS011269: posterior fornix
SRS011271: stool

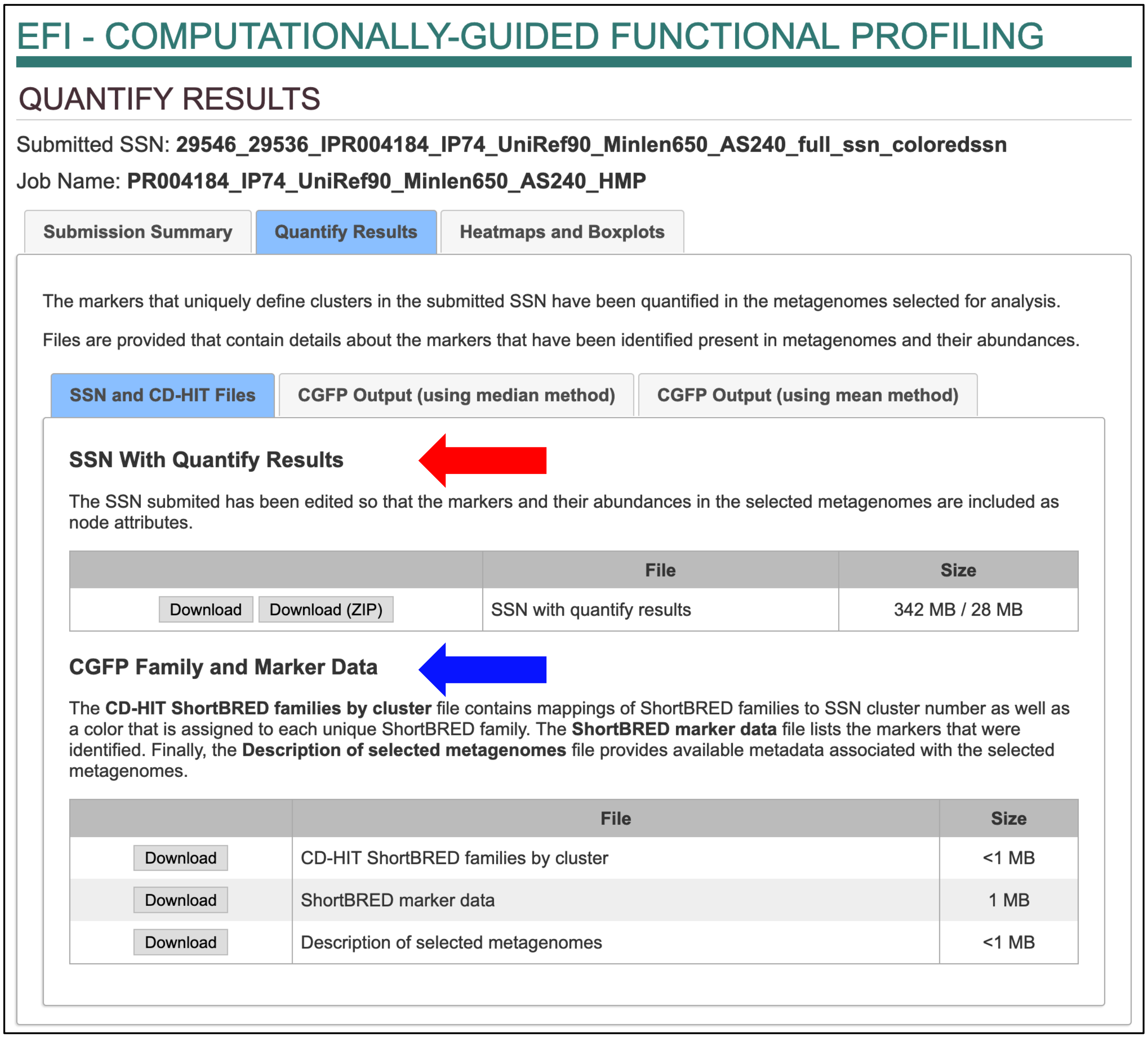
This page is available when quantification of marker abundance is completed.

“Submission Summary” tab. The “Submission Summary Table” provides the same job parameters provided in the “**Submission Summary Table**” on “**Marker Computation Results**” page plus the number of consequence sequences with metagenome hits that can be downloaded as a text file for future reference. This tab also provides a list of the metagenomes that were used for abundance quantification.

Metagenome Abundance Quantitation. According to the ShortBRED protocol, the metagenome abundance of a protein (and the cluster or singleton with which it is associated) is reported as the abundance of the median marker for the consensus sequence in the ShortBRED family (assuming that multiple markers are identified; the markers are arranged from N- to C-terminal in the consensus sequence); the default abundance coloring in the heat maps use this protocol. Abundance results also are calculated using the mean of the abundances determined for all markers in the consensus sequence. The user needs to be aware of the difference: the median protocol is more reliable, depending on the number of markers in a consensus sequence and the size/depth of the metagenome datasets, because it assumes “uniform” abundance across the consensus sequence; the mean protocol is more robust in identifying consensus sequences that have hits in the metagenome datasets. The abundances for the markers for the ShortBRED families contained in each SSN cluster are combined to quantitate SSN cluster abundance.

The metagenome abundances depicted in the heat maps and summarized in the files for download are normalized by average genome size (AGS) and range from one gene copy per metagenome to <0.0001 gene per metagenome. The actual dynamic range of the abundances is dependent on the depth of the metagenome dataset sequencing.

“Quantify Results” tab. This tab provides access (via three tabs) to an updated color SSN as well as text files that provide the marker abundance data.



EFI - COMPUTATIONALLY-GUIDED FUNCTIONAL PROFILING

QUANTIFY RESULTS

Submitted SSN: **29546_29536_IPR004184_IP74_UniRef90_Minlen650_AS240_full_ssn_coloredssn**

Job Name: **PR004184_IP74_UniRef90_Minlen650_AS240_HMP**

The markers that uniquely define clusters in the submitted SSN have been quantified in the metagenomes selected for analysis.

Files are provided that contain details about the markers that have been identified present in metagenomes and their abundances.

SSN and CD-HIT Files
CGFP Output (using median method)
CGFP Output (using mean method)

SSN With Quantify Results

The SSN submited has been edited so that the markers and their abundances in the selected metagenomes are included as node attributes.

	File	Size
Download Download (ZIP)	SSN with quantify results	342 MB / 28 MB

CGFP Family and Marker Data

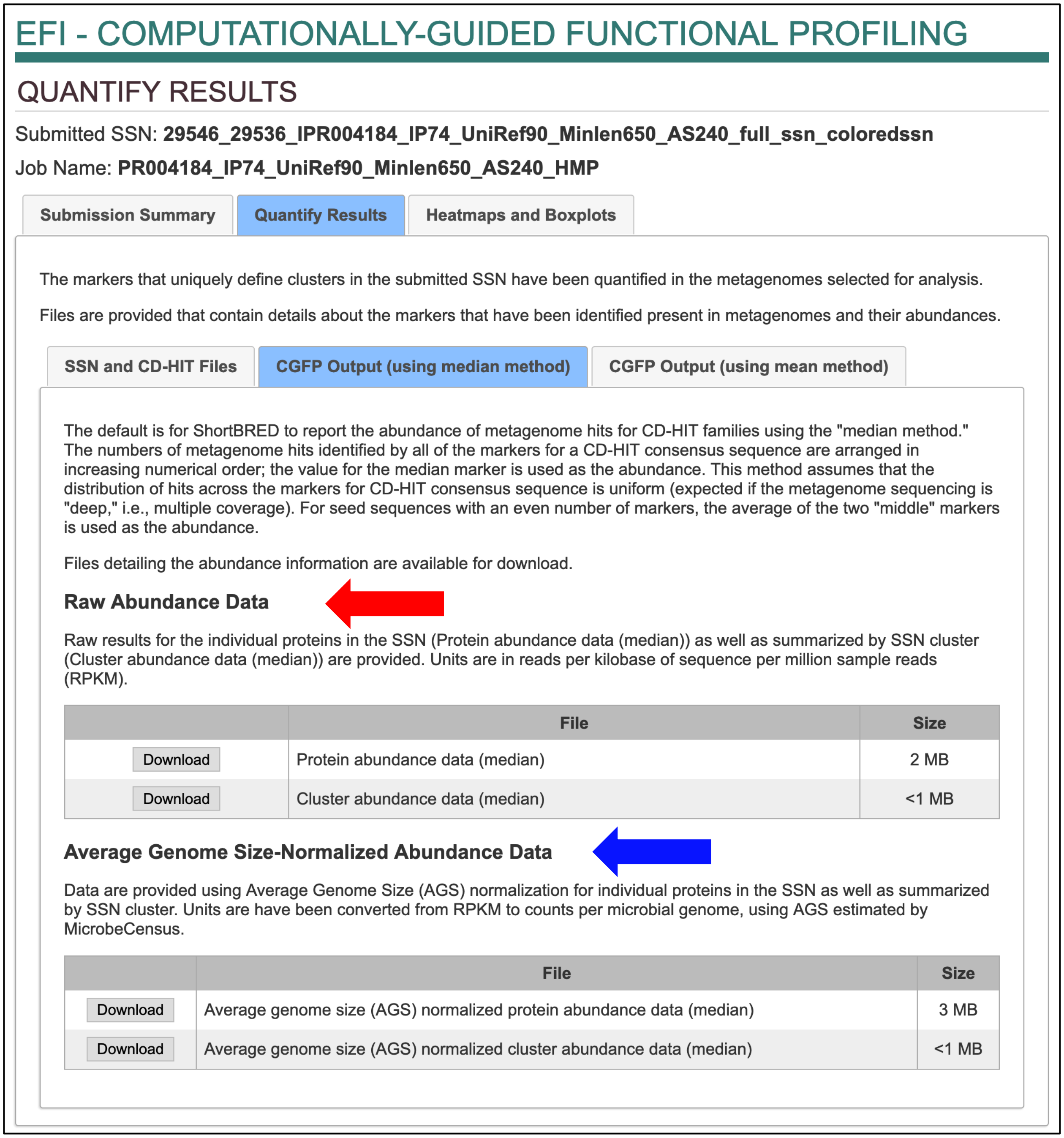
The **CD-HIT ShortBRED families by cluster** file contains mappings of ShortBRED families to SSN cluster number as well as a color that is assigned to each unique ShortBRED family. The **ShortBRED marker data** file lists the markers that were identified. Finally, the **Description of selected metagenomes** file provides available metadata associated with the selected metagenomes.

	File	Size
Download	CD-HIT ShortBRED families by cluster	<1 MB
Download	ShortBRED marker data	1 MB
Download	Description of selected metagenomes	<1 MB

“SSN and CD-HIT Files” tab. The “**SSN with Quantify Results**” panel (red arrow) provides a further updated version of the input color SSN to which node attributes are added for “Metagenomes Identified by Markers” [to those (meta)nodes that contain seed sequences for ShortBRED families] and “Metagenomes Identified by CD-HIT Family” [to those (meta)nodes that contribute sequences to the ShortBRED family] is available for download (“SSN with quantify results”). The (meta)nodes with entries in these node attributes can be selected with the Cytoscape Select panel so that the SSN (meta)nodes, clusters, and singletons with metagenome hits can be identified. The additional node attributes in the SSN generated by the metagenome abundance quantification step of EFI-CGFP are listed in Table S6.

In the “**CGFP Family and Marker Data**” panel (blue arrow), text files are available that define the membership of the ShortBRED families (“**CD-HIT ShortBRED families by clusters**”), provide the sequences of the markers determined for the ShortBRED family consensus sequences (“**ShortBRED marker data**”), and provide the metagenomes that were selected for quantification (“**Description of selected metagenomes**”).

"CGFP Output (using median method)" and "CGFP Output (using mean method)" tabs. These tabs provide text files that detail protein and cluster metagenome abundances [both "**Raw Abundance Data**" (red arrow) and "**Average Genome Size-Normalized Abundance Data**" (blue arrow)] that were calculated using both the median and mean methods. The cluster and singleton numbers in the files are those in the input colored SSN.



EFI - COMPUTATIONALLY-GUIDED FUNCTIONAL PROFILING

QUANTIFY RESULTS

Submitted SSN: **29546_29536_IPR004184_IP74_UniRef90_Minlen650_AS240_full_ssn_coloredssn**

Job Name: **PR004184_IP74_UniRef90_Minlen650_AS240_HMP**

The markers that uniquely define clusters in the submitted SSN have been quantified in the metagenomes selected for analysis.

Files are provided that contain details about the markers that have been identified present in metagenomes and their abundances.

SSN and CD-HIT Files
CGFP Output (using median method)
CGFP Output (using mean method)

The default is for ShortBRED to report the abundance of metagenome hits for CD-HIT families using the "median method." The numbers of metagenome hits identified by all of the markers for a CD-HIT consensus sequence are arranged in increasing numerical order; the value for the median marker is used as the abundance. This method assumes that the distribution of hits across the markers for CD-HIT consensus sequence is uniform (expected if the metagenome sequencing is "deep," i.e., multiple coverage). For seed sequences with an even number of markers, the average of the two "middle" markers is used as the abundance.

Files detailing the abundance information are available for download.

Raw Abundance Data

Raw results for the individual proteins in the SSN (Protein abundance data (median)) as well as summarized by SSN cluster (Cluster abundance data (median)) are provided. Units are in reads per kilobase of sequence per million sample reads (RPKM).

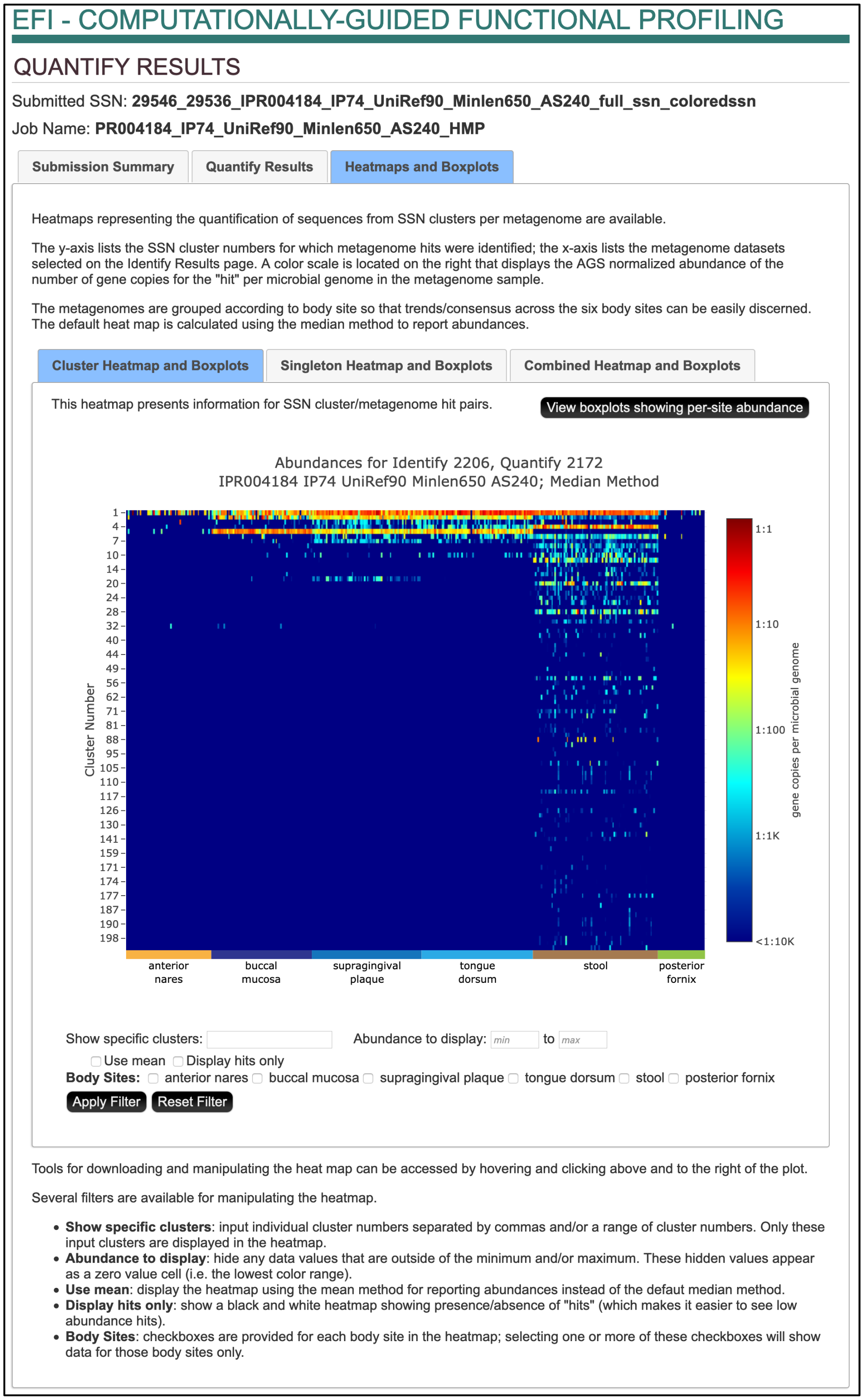
	File	Size
Download	Protein abundance data (median)	2 MB
Download	Cluster abundance data (median)	<1 MB

Average Genome Size-Normalized Abundance Data

Data are provided using Average Genome Size (AGS) normalization for individual proteins in the SSN as well as summarized by SSN cluster. Units are have been converted from RPKM to counts per microbial genome, using AGS estimated by MicrobeCensus.

	File	Size
Download	Average genome size (AGS) normalized protein abundance data (median)	3 MB
Download	Average genome size (AGS) normalized cluster abundance data (median)	<1 MB

"Heatmaps and Boxplots" tab. In this tab, three additional tabs are provided: "**Cluster Heatmap and Boxplots**" for SSN clusters, "**Singleton Heatmap and Boxplots**" for SSN singletons, and "**Combined Heatmap and Boxplots**" for both SSN clusters and singletons.



EFI - COMPUTATIONALLY-GUIDED FUNCTIONAL PROFILING

QUANTIFY RESULTS

Submitted SSN: **29546_29536_IPR004184_IP74_UniRef90_Minlen650_AS240_full_ssn_coloredssn**

Job Name: **PR004184_IP74_UniRef90_Minlen650_AS240_HMP**

Heatmaps representing the quantification of sequences from SSN clusters per metagenome are available.

The y-axis lists the SSN cluster numbers for which metagenome hits were identified; the x-axis lists the metagenome datasets selected on the Identify Results page. A color scale is located on the right that displays the AGS normalized abundance of the number of gene copies for the "hit" per microbial genome in the metagenome sample.

The metagenomes are grouped according to body site so that trends/consensus across the six body sites can be easily discerned. The default heat map is calculated using the median method to report abundances.

Cluster Heatmap and Boxplots
Singleton Heatmap and Boxplots
Combined Heatmap and Boxplots

This heatmap presents information for SSN cluster/metagenome hit pairs.

Show specific clusters: ▭ Abundance to display: min to max

□ Use mean □ Display hits only

Body Sites: □ anterior nares □ buccal mucosa □ supragingival plaque □ tongue dorsum □ stool □ posterior fornix

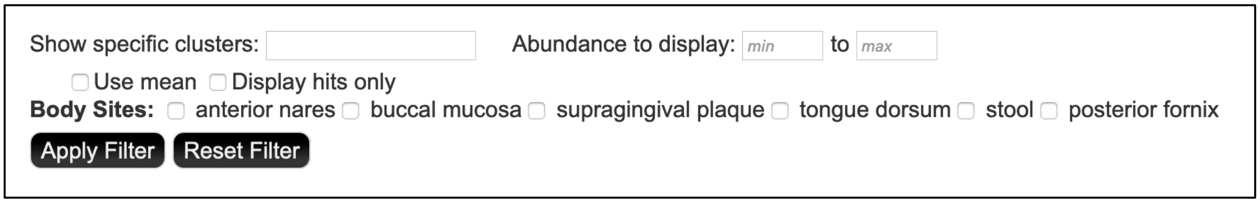
Tools for downloading and manipulating the heat map can be accessed by hovering and clicking above and to the right of the plot.

Several filters are available for manipulating the heatmap.

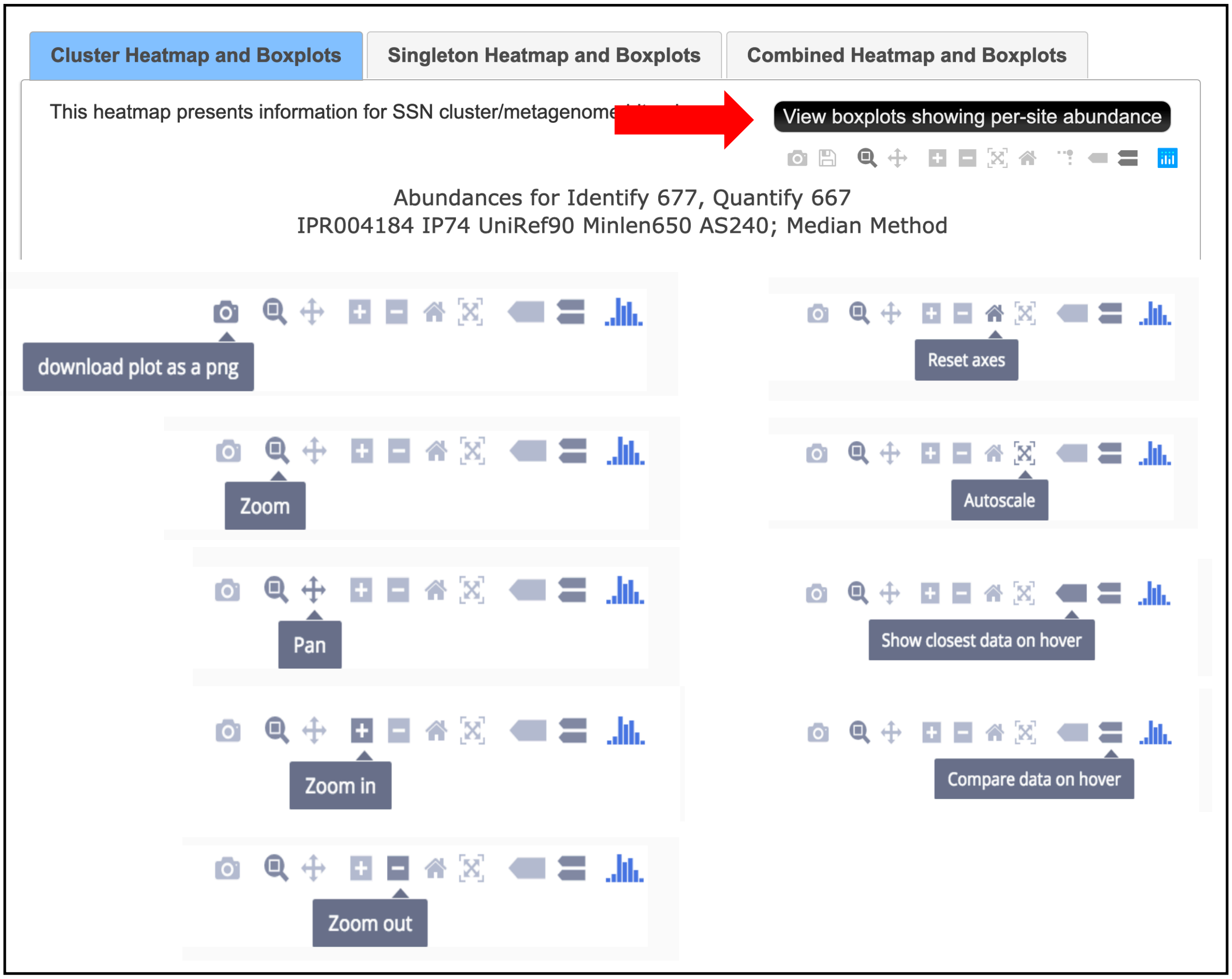
- **Show specific clusters**: input individual cluster numbers separated by commas and/or a range of cluster numbers. Only these input clusters are displayed in the heatmap.
- **Abundance to display**: hide any data values that are outside of the minimum and/or maximum. These hidden values appear as a zero value cell (i.e. the lowest color range).
- **Use mean**: display the heatmap using the mean method for reporting abundances instead of the defaut median method.
- **Display hits only**: show a black and white heatmap showing presence/absence of "hits" (which makes it easier to see low abundance hits).
- **Body Sites**: checkboxes are provided for each body site in the heatmap; selecting one or more of these checkboxes will show data for those body sites only.

For each heatmap and boxplot tab, the default display is the heatmap that summarizes the abundance data. Heatmaps and boxplots (*vide infra*) are displayed using the Plotly graphics library[15]. The heat maps are interactive, allowing the user to select a range of clusters/singletons or specific clusters/singletons; the cluster/singleton numbers are those assigned in the input color SSN, with "S" proceeding singleton cluster numbers. By hovering the pointer over the heatmap, the user can identify the dataset associated with each metagenome hit and its abundance.

The user can use the panel below the heatmap to select specific body sites for display, a range of abundances, and whether the displayed metagenome abundance is calculated using the mean or median method (*vide supra*). The "**Display hits only**" box provides a black and white "heatmap" so that low abundance "hits" can be more readily distinguished from the background.


Show specific clusters:
Abundance to display: min to max
□ Use mean □ Display hits only
Body Sites: □ anterior nares □ buccal mucosa □ supragingival plaque □ tongue dorsum □ stool □ posterior fornix
Apply Filter
Reset Filter

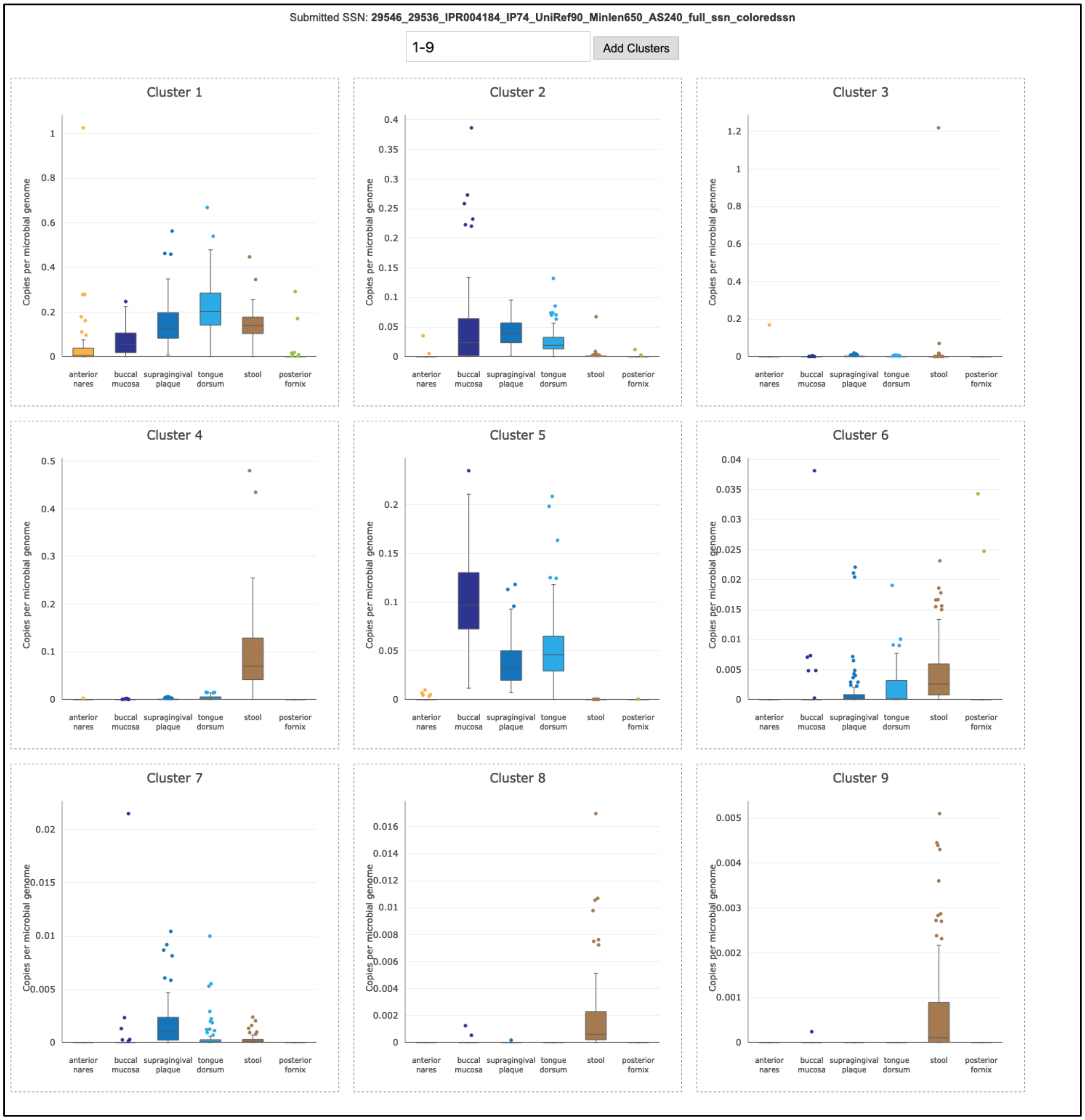
By hovering the pointer over the upper right-hand corner of the heatmap panel, icons appear that allow the heatmap to be downloaded as a png file and provide options for altering the display.


Cluster Heatmap and Boxplots
Singleton Heatmap and Boxplots
Combined Heatmap and Boxplots
View boxplots showing per-site abundance
download plot as a png
Reset axes
Zoom
Autoscale
Pan
Show closest data on hover
Zoom in
Compare data on hover
Zoom out

This heatmap presents information for SSN cluster/metagenome

Abundances for Identify 677, Quantify 667
IPR004184 IP74 UniRef90 Minlen650 AS240; Median Method

The **"View boxplots showing per-site abundance"** button located in the upper right corner of each heatmap tab (red arrow) provides access to boxplots that provide quantitative analyses of the median abundance data for each cluster or singleton [using the abundance data available for download in the "**CGFP Output (using median method)**" tab in the "**Quantify Results**" tab of this page]. A new window (can be expanded horizontally and vertically) opens with a box to enter a list and/or range of cluster and/or singleton numbers.



Submitted SSN: **29546_29536_IPR004184_IP74_UniRef90_Minlen650_AS240_full_ssn_coloredssn**

1-9

For each selected cluster/singleton, a box plot is displayed showing the metagenome abundances for each body site; by hovering the pointer over the entry for each body site, the minimum, maximum, first quartile, third quartile, and mean abundance values are displayed. By hovering the pointer over the upper right-hand corner of the boxplot window, the icons are displayed and allow the box plots to be downloaded as well as provide several options for altering the display of the box plot. All of the abundance data displayed in the heatmaps and boxplots are available in the tables that can be downloaded from the "**CGFP Output**" tabs if the user would like to perform additional statistical/quantitative analyses.

References

[1] Atkinson, H. J., Morris, J. H., Ferrin, T. E., and Babbitt, P. C. (2009) Using sequence similarity networks for visualization of relationships across diverse protein superfamilies, *PLoS One 4*, e4345.

[2] Gerlt, J. A., Bouvier, J. T., Davidson, D. B., Imker, H. J., Sadkhin, B., Slater, D. R., and Whalen, K. L. (2015) Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): A web tool for generating protein sequence similarity networks, *Biochim Biophys Acta 1854*, 1019-1037.

[3] Brown, S. D., and Babbitt, P. C. (2012) Inference of functional properties from large-scale analysis of enzyme superfamilies, *J Biol Chem 287*, 35-42.

[4] Copp, J. N., Akiva, E., Babbitt, P. C., and Tokuriki, N. (2018) Revealing Unexplored Sequence-Function Space Using Sequence Similarity Networks, *Biochemistry 57*, 4651-4662.

[5] Copp, J. N., Anderson, D. W., Akiva, E., Babbitt, P. C., and Tokuriki, N. (2019) Exploring the sequence, function, and evolutionary space of protein superfamilies using sequence similarity networks and phylogenetic reconstructions, *Methods Enzymol 620*, 315-347.

[6] Brown, S. D., and Babbitt, P. C. (2014) New insights about enzyme evolution from large scale studies of sequence and structure relationships, *J Biol Chem 289*, 30221-30228.

[7] Pandya, C., Brown, S., Pieper, U., Sali, A., Dunaway-Mariano, D., Babbitt, P. C., Xia, Y., and Allen, K. N. (2013) Consequences of domain insertion on sequence-structure divergence in a superfold, *Proc Natl Acad Sci U S A 110*, E3381-3387.

[8] Pandya, C., Dunaway-Mariano, D., Xia, Y., and Allen, K. N. (2014) Structure-guided approach for detecting large domain inserts in protein sequences as illustrated using the haloacid dehalogenase superfamily, *Proteins 82*, 1896-1906.

[9] Levin, B. J., Huang, Y. Y., Peck, S. C., Wei, Y., Martinez-Del Campo, A., Marks, J. A., Franzosa, E. A., Huttenhower, C., and Balskus, E. P. (2017) A prominent glycyl radical enzyme in human gut microbiomes metabolizes trans-4-hydroxy-l-proline, *Science 355*.

[10] Kaminski, J., Gibson, M. K., Franzosa, E. A., Segata, N., Dantas, G., and Huttenhower, C. (2015) High-Specificity Targeted Functional Profiling in Microbial Communities with ShortBRED, *PLoS Comput Biol 11*, e1004557.

[11] Li, W., and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, *Bioinformatics 22*, 1658-1659.

[12] Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data, *Bioinformatics 28*, 3150-3152.

[13] Buchfink, B., Xie, C., and Huson, D. H. (2015) Fast and sensitive protein alignment using DIAMOND, *Nat Methods 12*, 59-60.

[14] Edgar, R. C. (2010) Search and clustering orders of magnitude faster than BLAST, *Bioinformatics 26*, 2460-2461.