EFI-SSN and Sequence-Based Functional Prediction in Enzyme Discovery

BlueBio4Future Short Course in Bioinformatics

18th March 2025

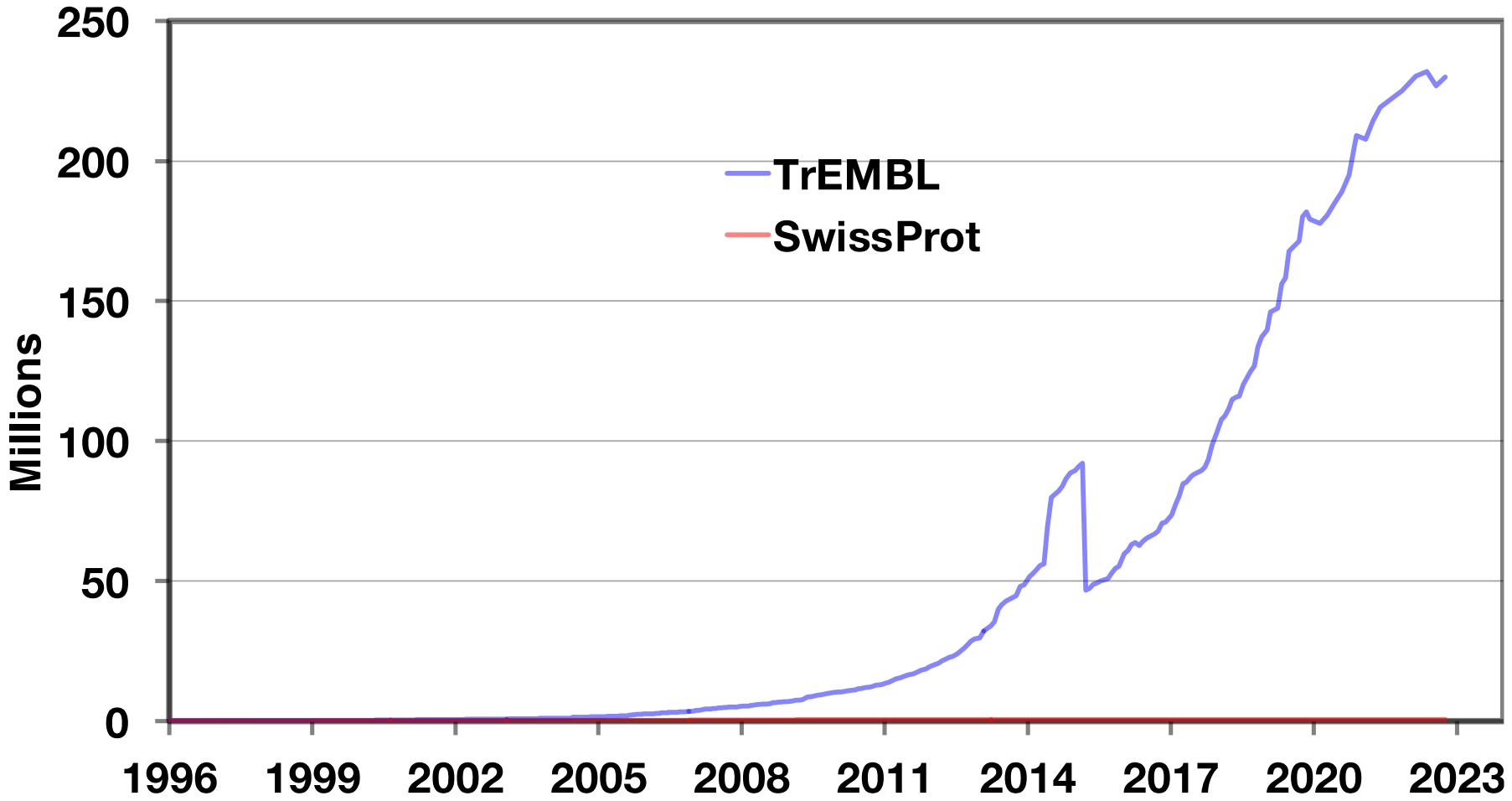
Raquel Castelo Branco

Background

The number of protein sequences is “exploding” !

UniProtKB/TrEMBL database contains **computationally annotated** 229,928,140 entries (Release October-2022).

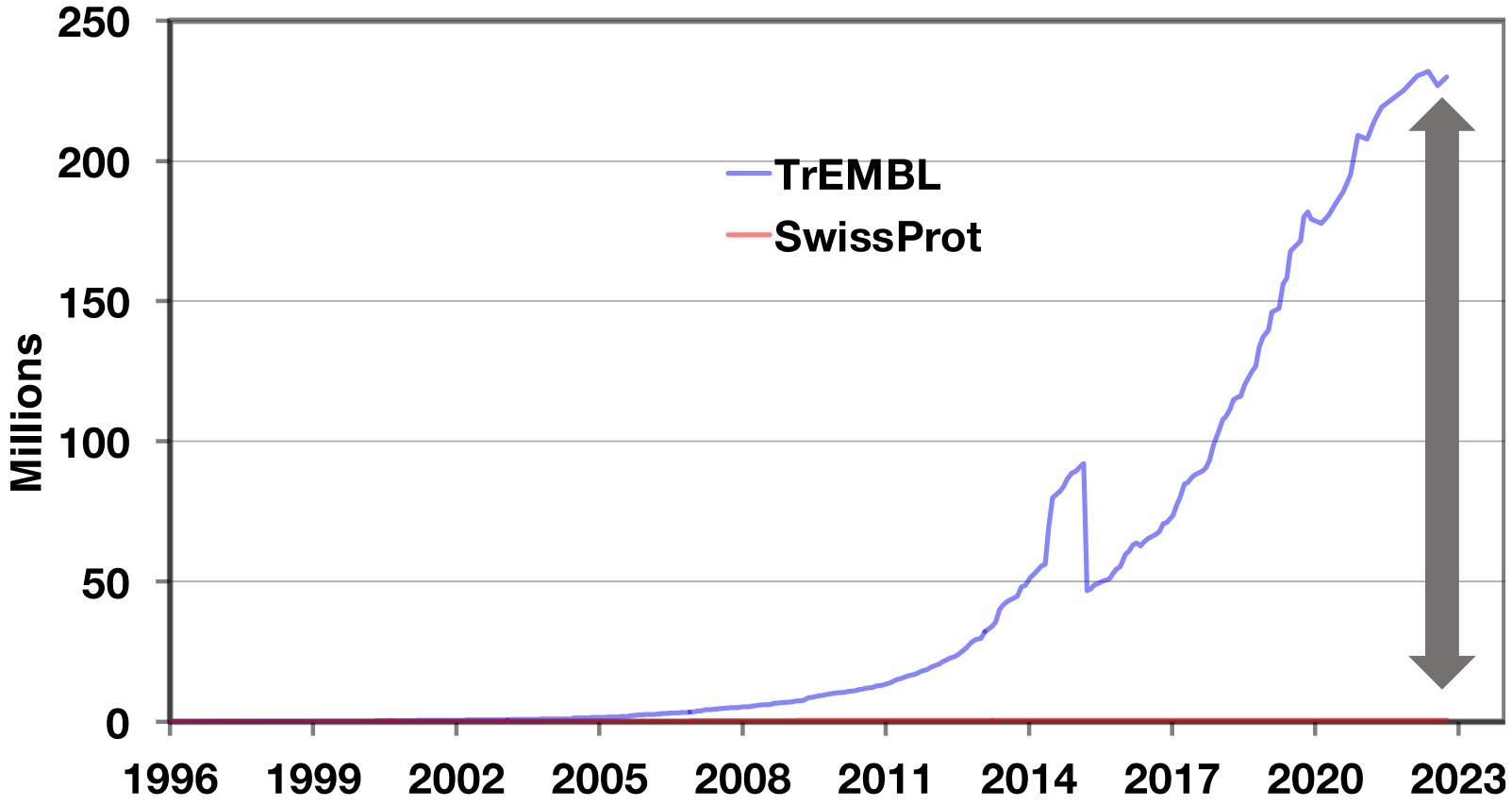
UniProtKB/SwissProt database contains **manually curated** 568,363 entries (Release October-2022).



Background

BlueBio Future

Perhaps 50% have unknown or uncertain functions
How do we solve this problem ?

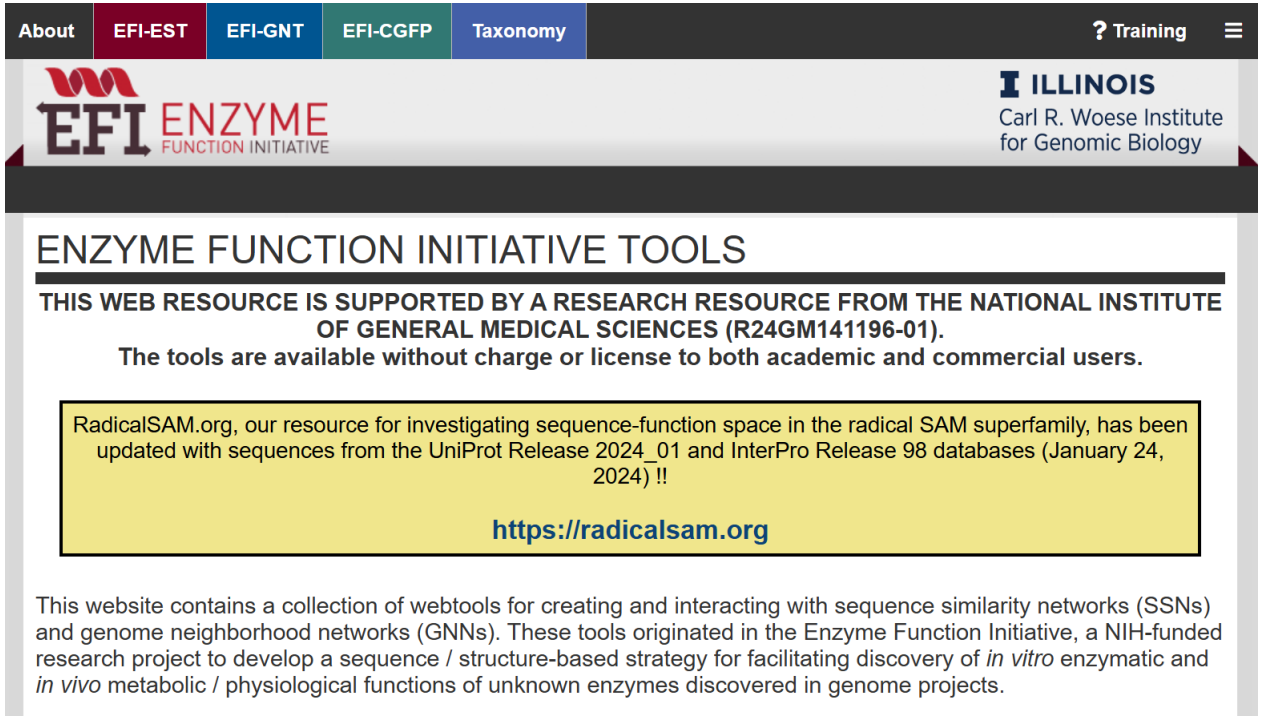


Requires large-scale bioinformatics, computational and experimental tools and strategies to use them

Enzyme Function Initiative-Enzyme Similarity Tool

- A web tool for **visualizing protein sequence relationships** as a network – EFI-EST.
- Helps in **functional annotation**, **evolutionary studies**, and **protein clustering**.

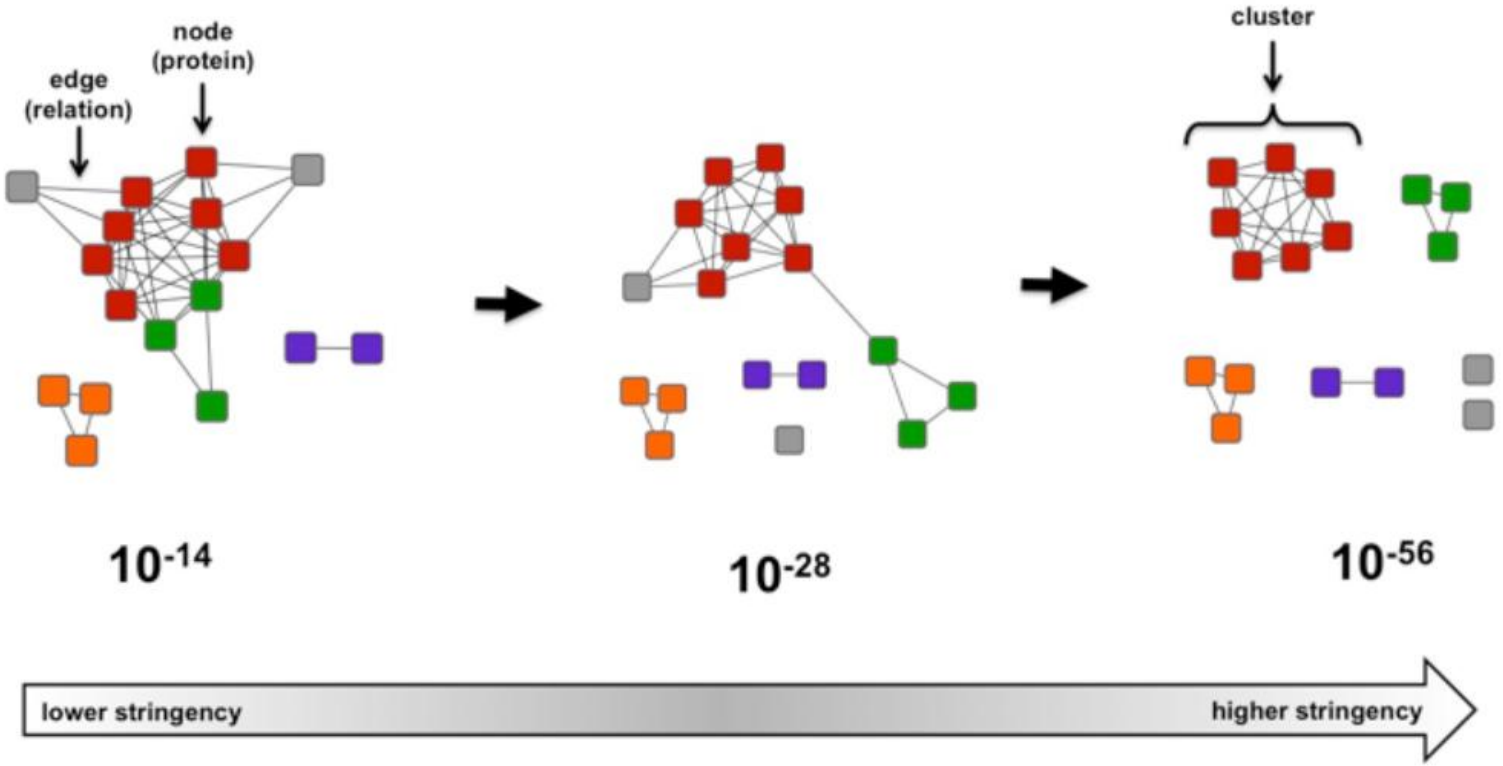
https://efi.igb.illinois.edu/



- ✓ **SSNs show global relationships** within protein families.
- ✓ Useful for **metagenomics, drug discovery, enzyme function prediction**.

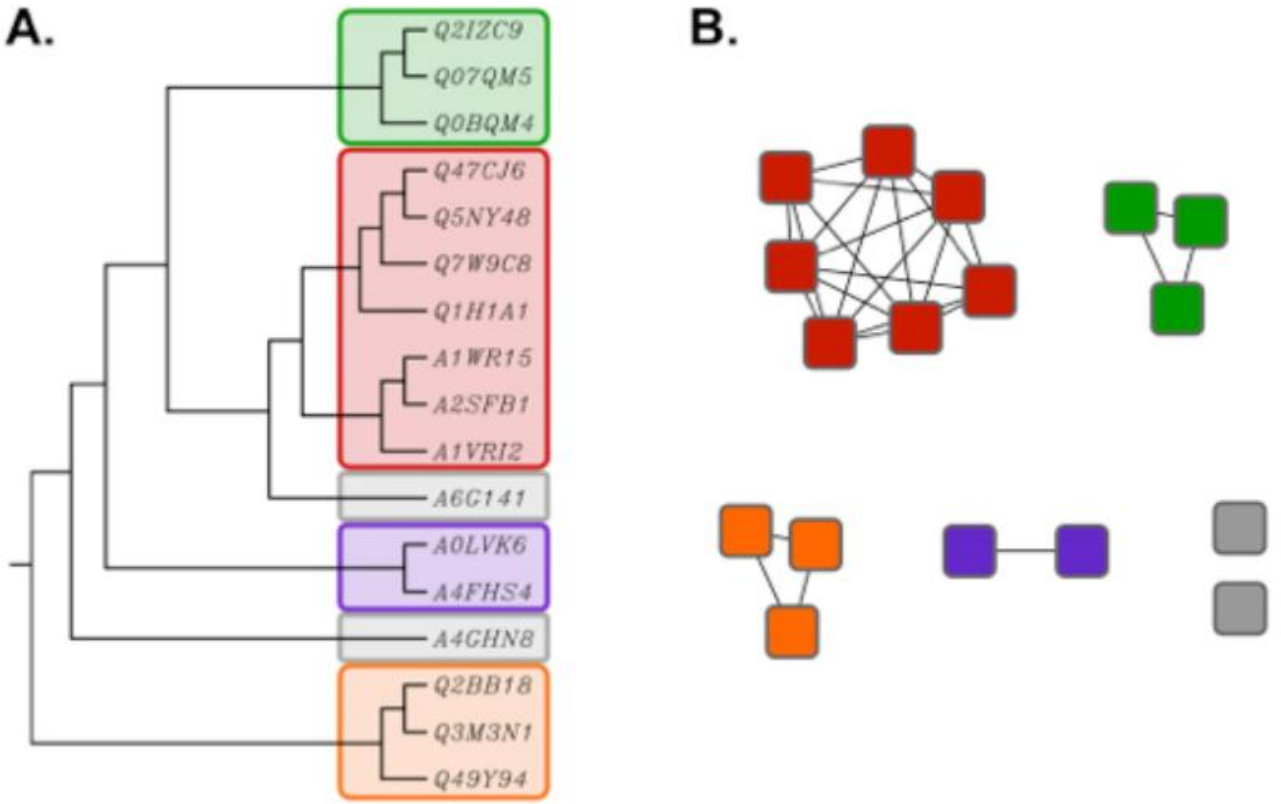
How Does EFI-SSN Work?

- **Nodes** = protein sequences.
- **Edges** = sequence similarity (BLAST E-value or % identity).
- Sequences **cluster based on functional or evolutionary relationships**.



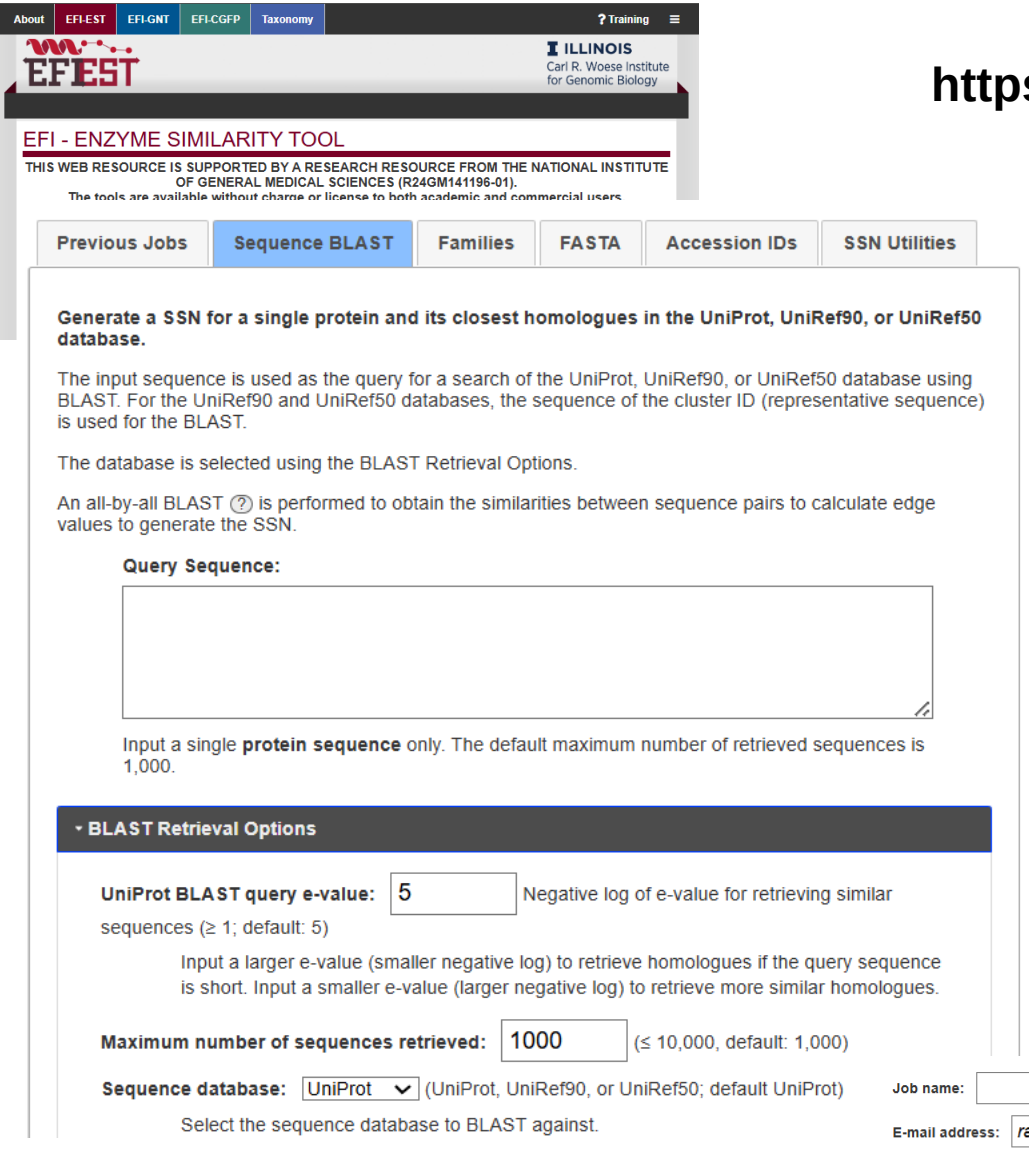
SSN vs Phylogenetic Trees

Although not as rigorous as traditional phylogenetic trees, SSNs typically display the same topology.



However, the advantage of SSNs over trees is that **large sequence sets** (e.g. many thousands of proteins) can be analyzed much **more quickly**, and visualized **easily** using Cytoscape.

SSN Input - General Considerations

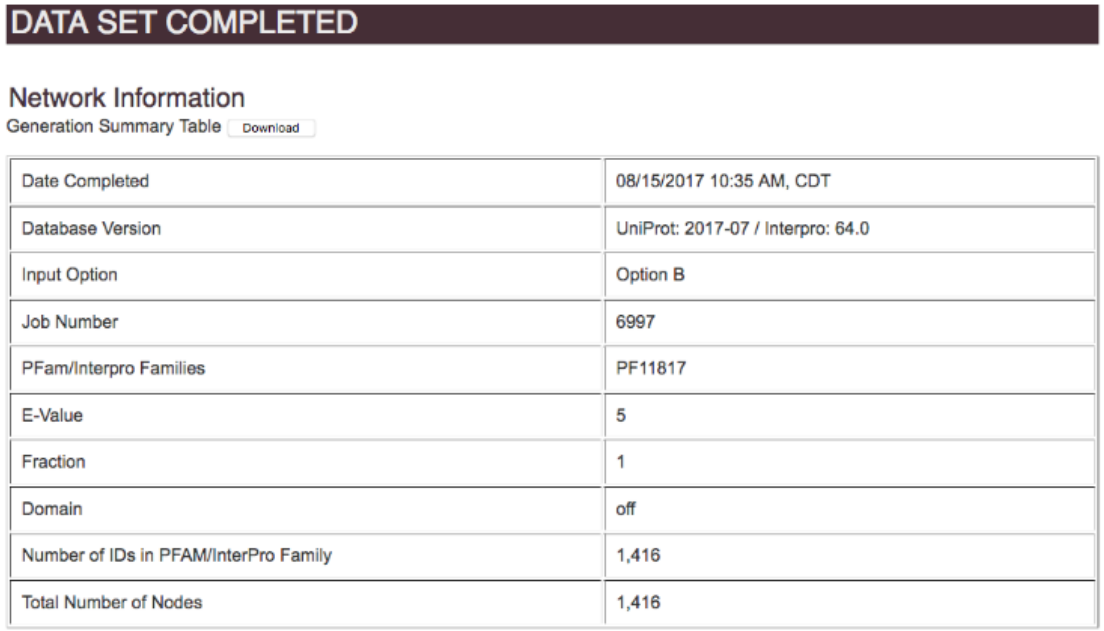


https://efi.igb.illinois.edu/

Four options for generating SSNs

- **Option A:** Homologues in UniProt to user-supplied sequences (collected by BLAST).
- **Option B:** User-supplied Pfam and/or InterPro families from UniProt, using full length sequences or domains.
- **Option C:** User-supplied FASTA file.
- **Option D:** User-supplied file of **UniProt** and/or **NCBI** IDs.

Data Set Complete Page

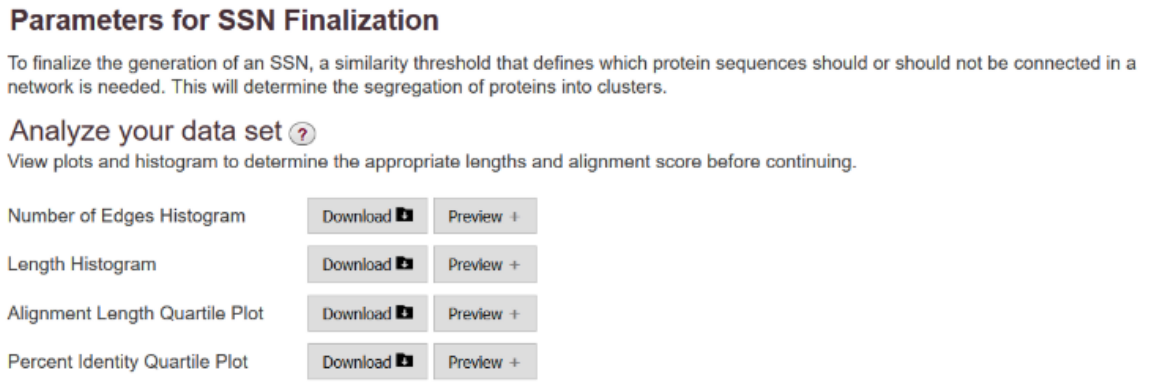


DATA SET COMPLETED

Network Information

Generation Summary Table Download

Date Completed	08/15/2017 10:35 AM, CDT
Database Version	UniProt: 2017-07 / Interpro: 64.0
Input Option	Option B
Job Number	6997
PFam/Interpro Families	PF11817
E-Value	5
Fraction	1
Domain	off
Number of IDs in PFAM/InterPro Family	1,416
Total Number of Nodes	1,416



Parameters for SSN Finalization

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

Analyze your data set

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

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

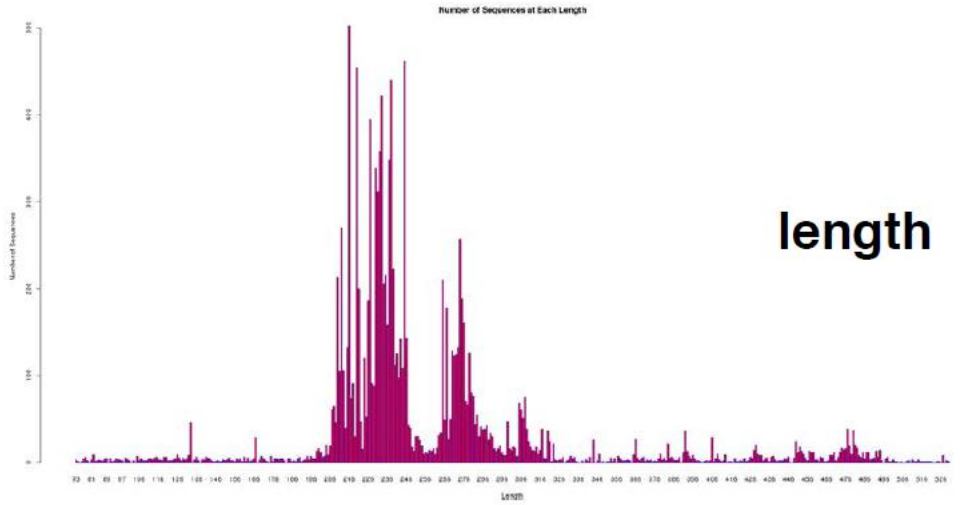
- The **information** needed for the generation of the dataset has been **processed** and the **similarity** between the sequences retrieved has been **calculated**.

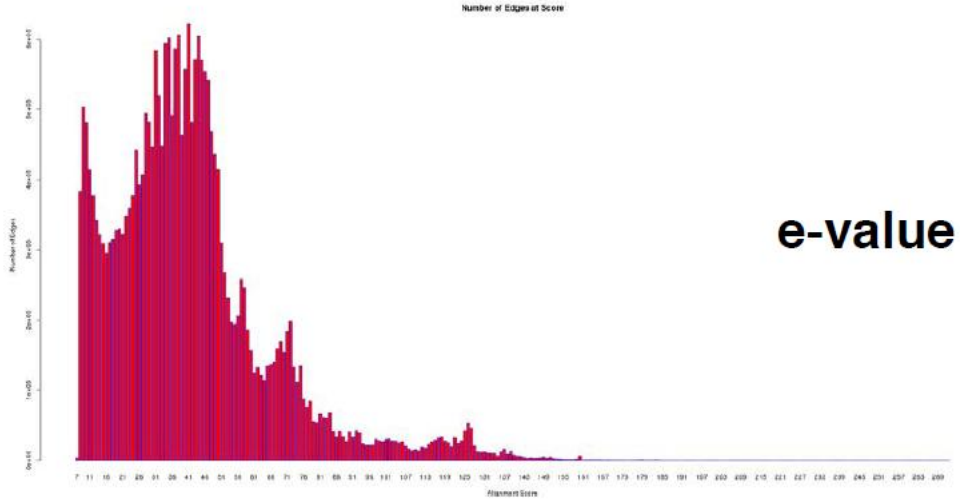
- Now the **Dataset Analysis** page provides a **summary about the input used**, and the returned calculations for the all-by-all BLAST.

- You **must interpret the provided information** in order to choose an **alignment score** that will be used for the **final step** of the **SSN generation**.

Data Set Complete Page

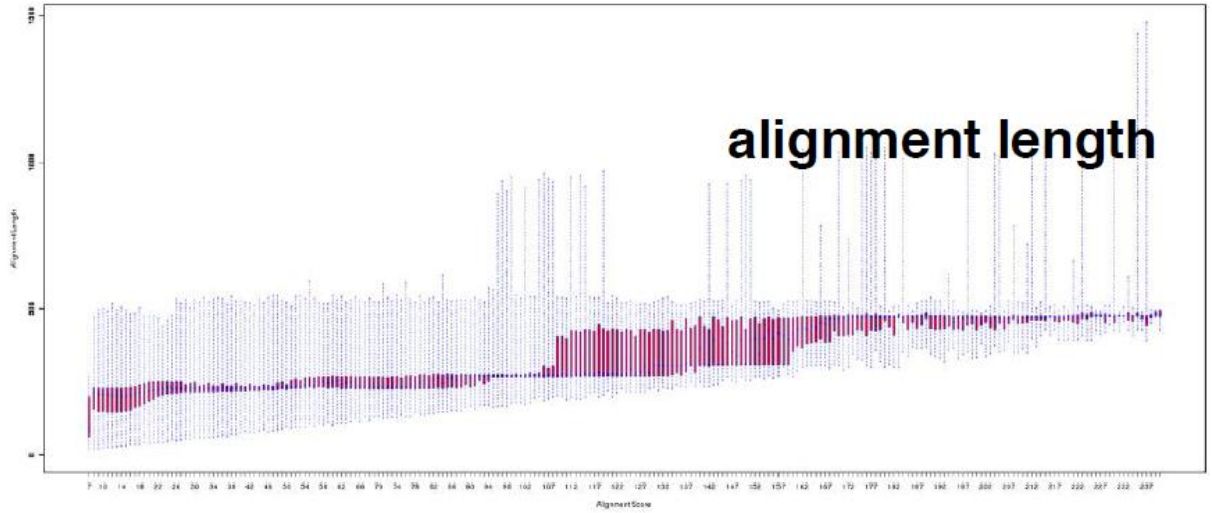
Two histograms: length and e-value

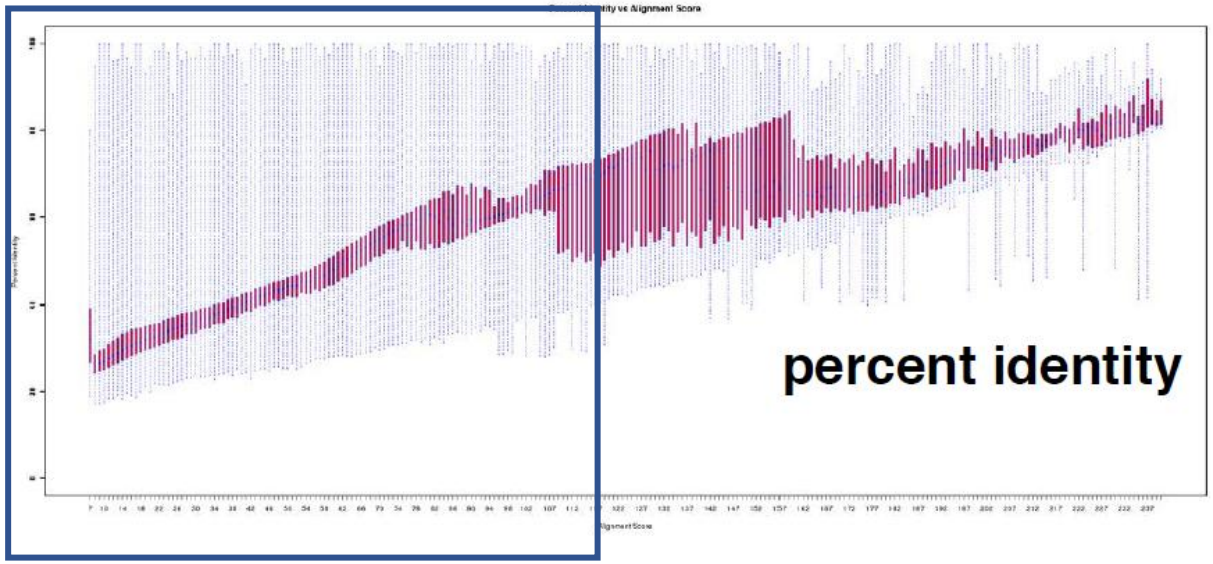




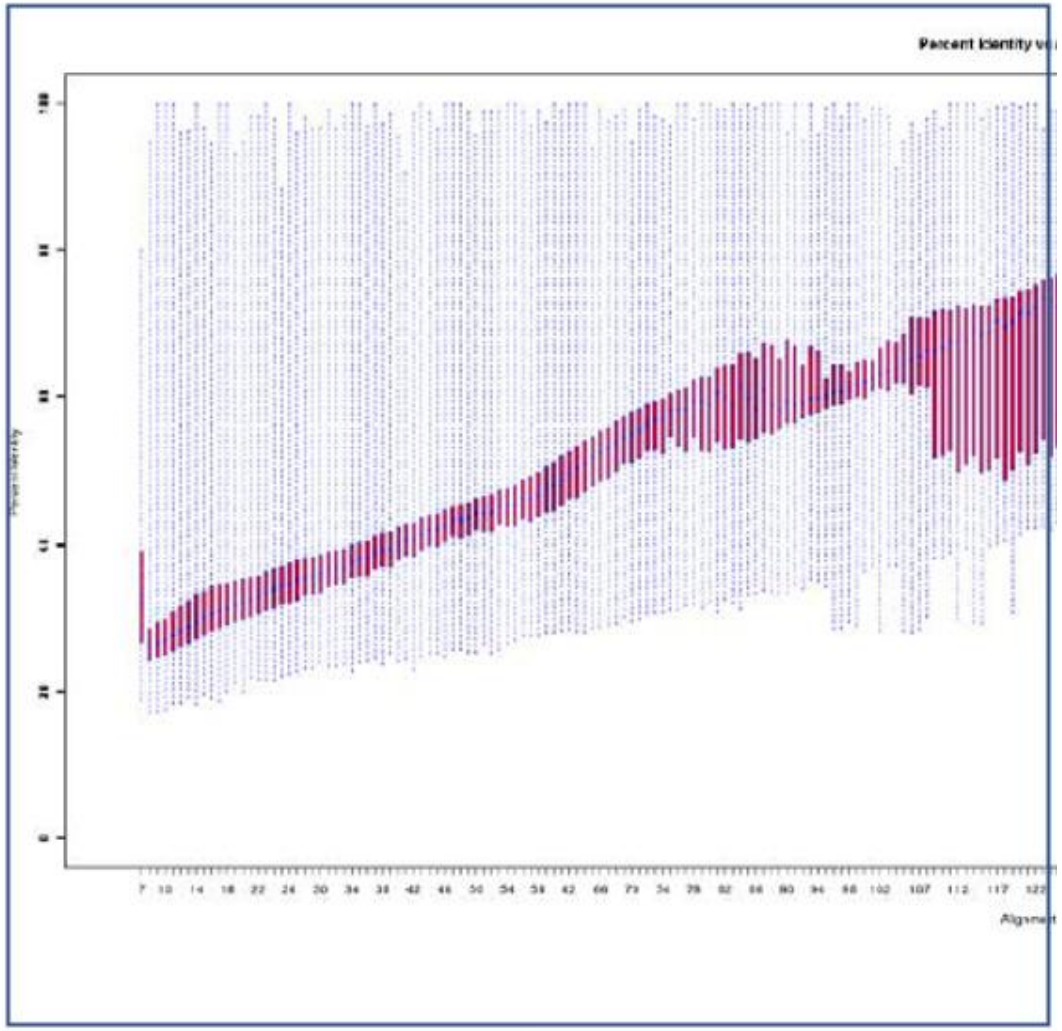
Data Set Complete Page

Quartile plots: alignment length and percent identity





Data Set Complete Page



Choosing an appropriate threshold

Recommended procedure:

Generate an initial SSN with a "low" alignment score so that isofunctional families are not separated. Although this alignment score will depend on the family, a useful "rule-of-thumb" is that isofunctional families often share >40% sequence identity.

Thus, it is recommend that the alignment score used to output the initial SSN should correspond to a lower sequence identity, e.g., 35%.

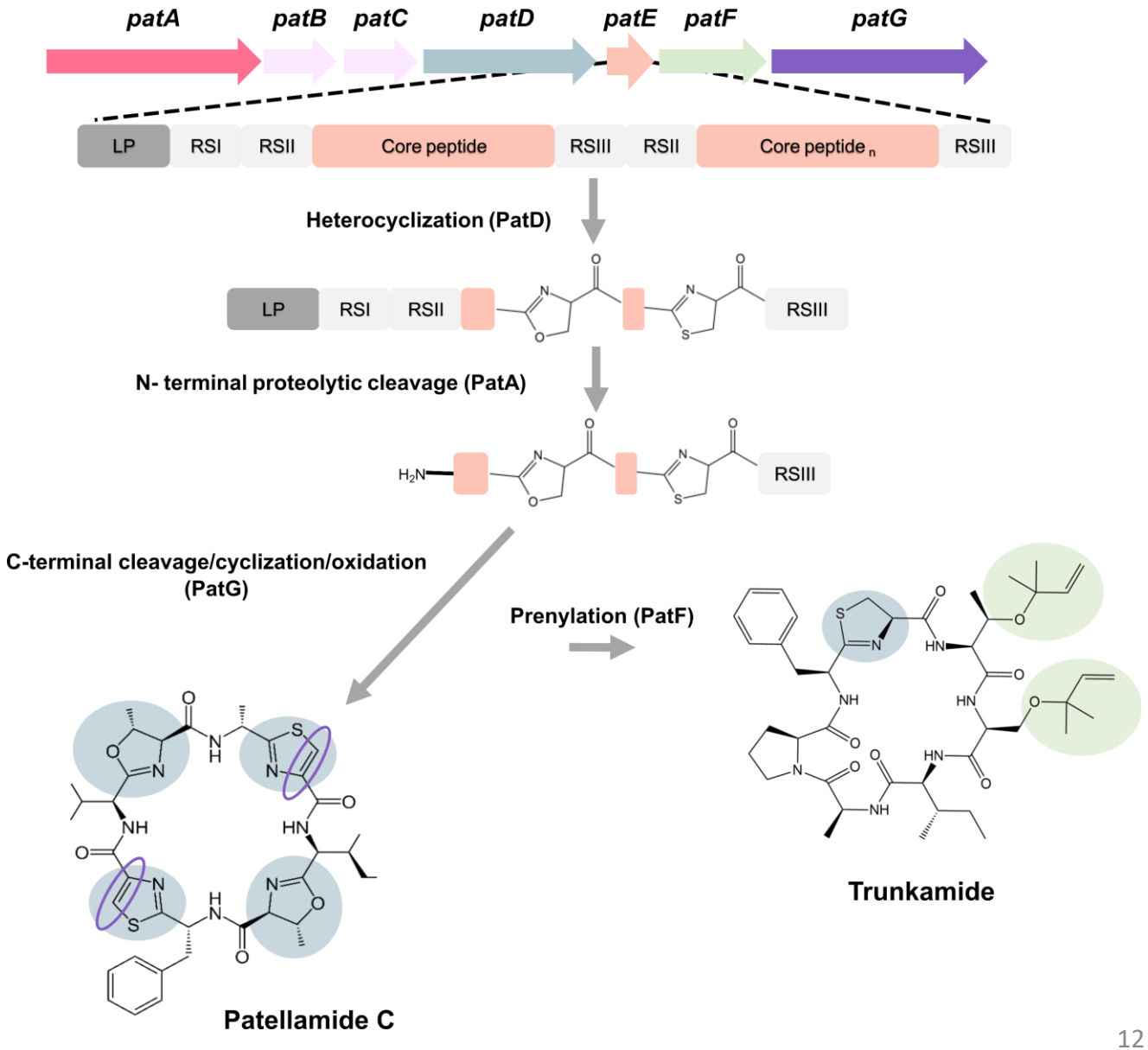
How to Use EFI-SSN – Case Study

Cyanobactins

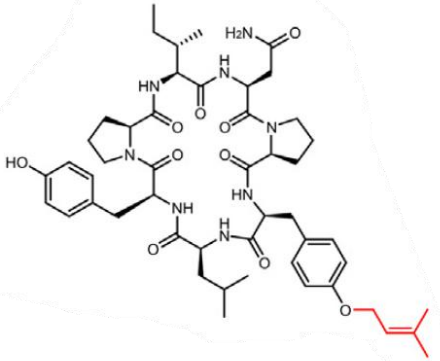
- Cyclic or linear RiPPs
- Produced by symbiotic and free-living cyanobacteria;

Several modifications

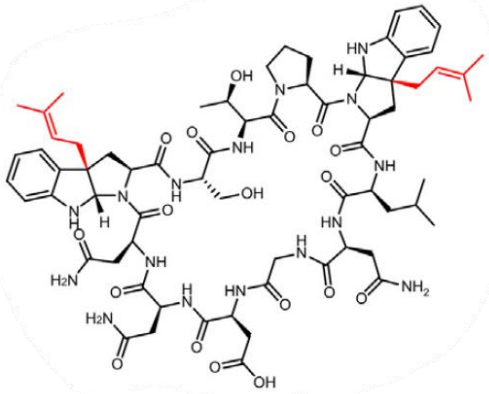
- N-to-C macrocyclization
- Heterocyclization of Ser, Thr and Cys
- Oxidation, **prenylation**, methylation, amidation



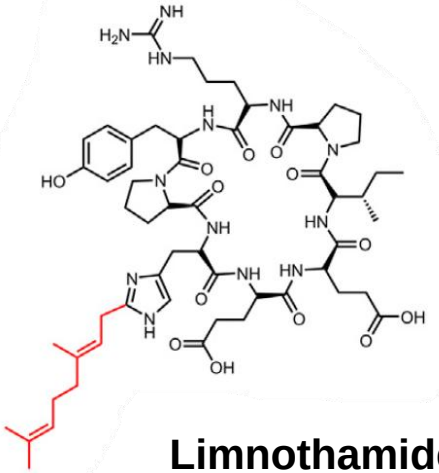
How to Use EFI-SSN – Case Study



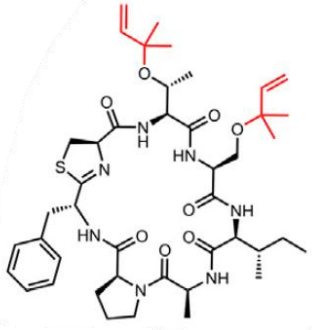
Prenylagaramide B
PagF (Tyr-*O*-Ptases)



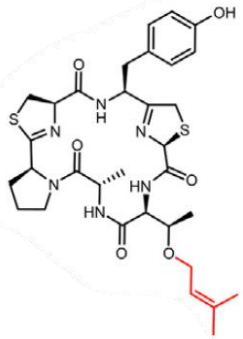
Kawaguchipeptin A
KgpF (Trp-*C*-Ptases)



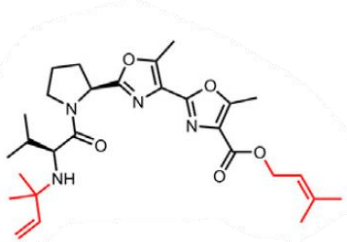
Limnothamide
LimF (His-*C*-Ptases)



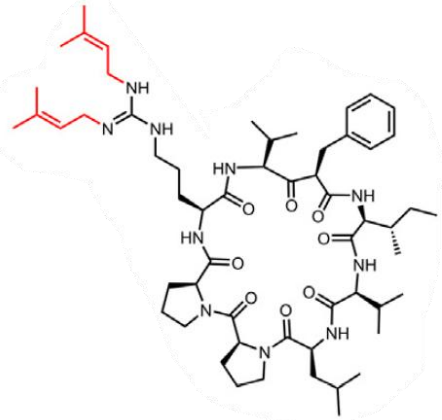
Trunkamide
TruF1 (Ser/Thr-*O*-Ptases)



Tolypamide
TolF (Ser/Thr-*O*-Ptases)



Muscoride A
MusF1/MusF2 (Terminus-*N*/*O*-Ptases)



Argicyclamide A
AgcF (Arg-*N*-Ptases)

How to Use EFI-SSN – Case Study

JOURNAL OF NATURAL PRODUCTS

Open Access

pubs.acs.org/jnp

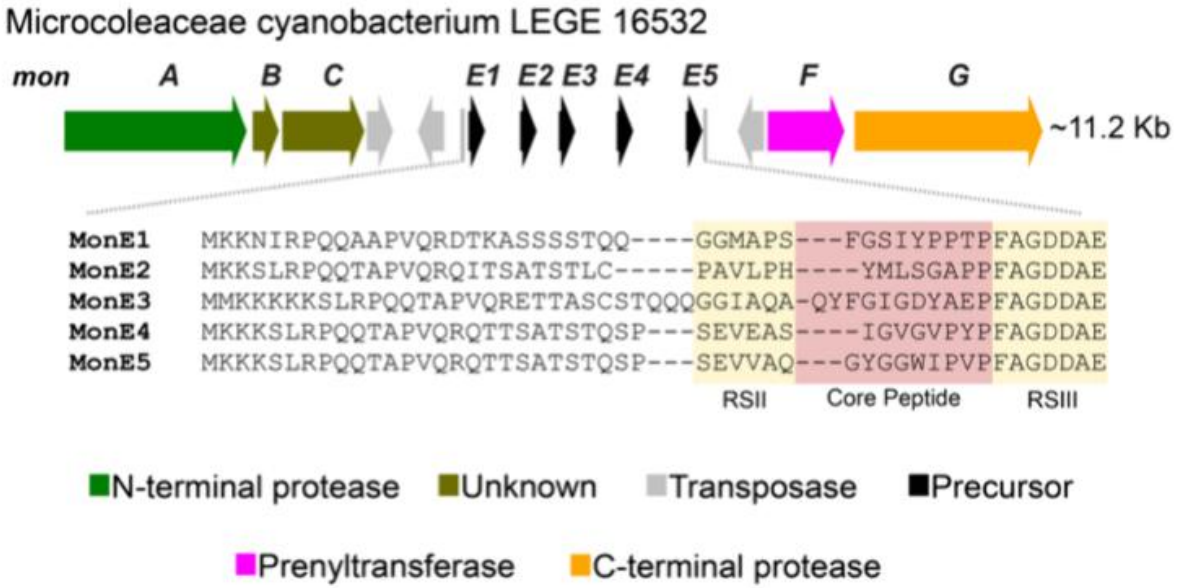
Article

Genome-informed Discovery of Monchicamides A–K: Cyanobactins from the Microcoleaceae Cyanobacterium LEGE 16532

Raquel Castelo-Branco,[Δ] João P. Pereira,[Δ] Sara Freitas, Marco Preto, Ana R. Vieira, João Morais, and Pedro N. Leão*

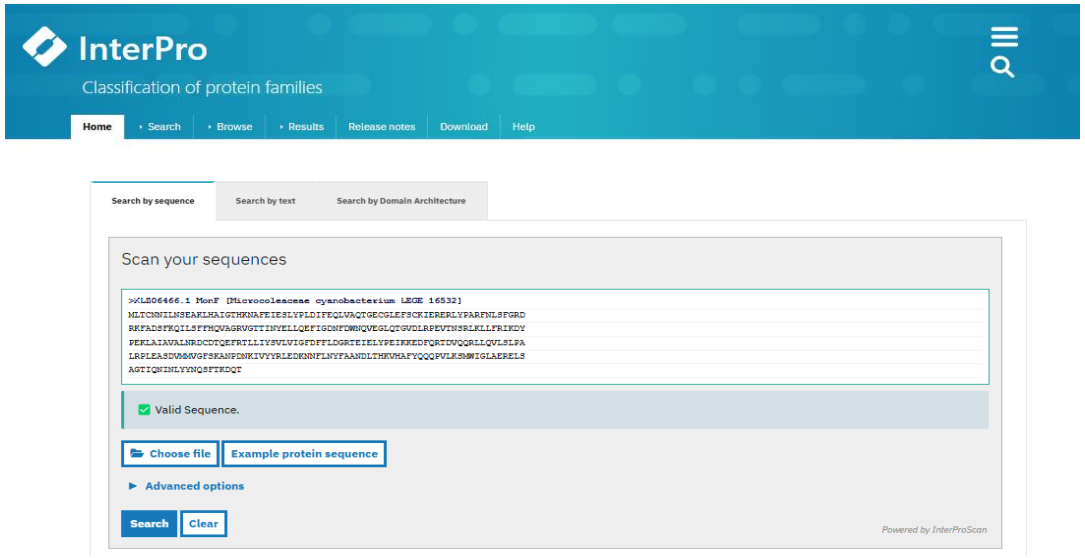
Cite This: *J. Nat. Prod.* 2025, 88, 86–93

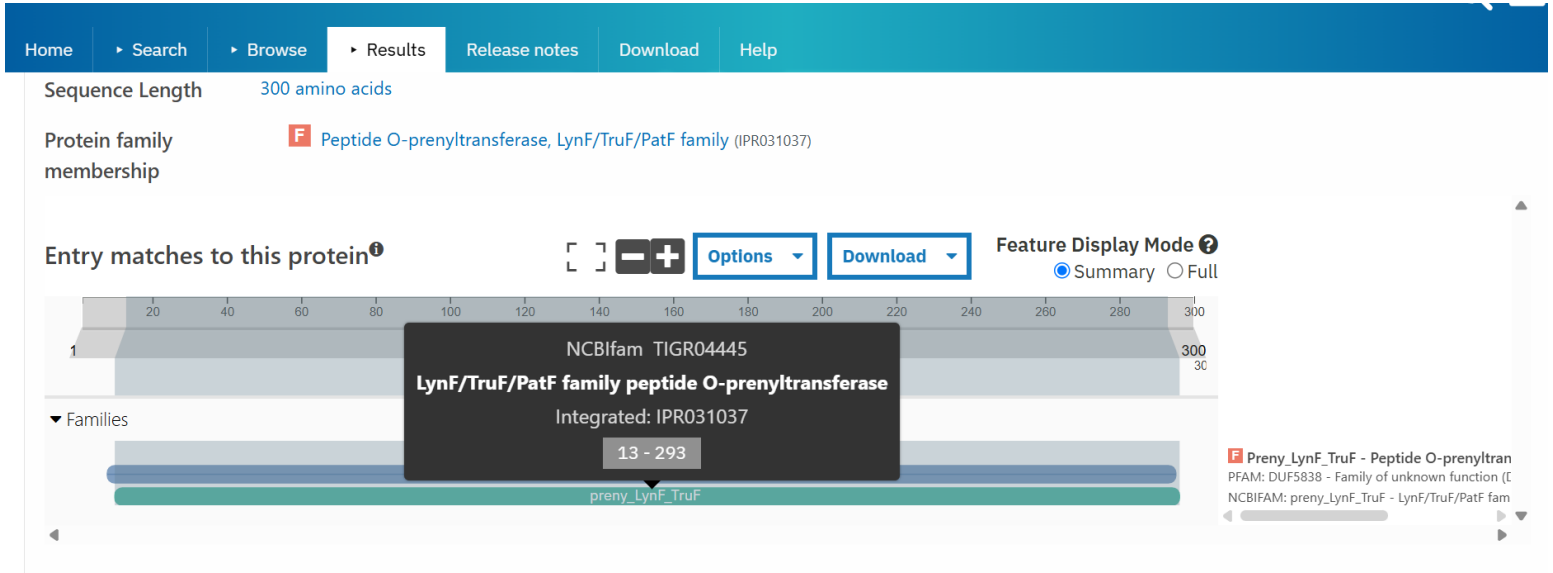
Read Online



How to Use EFI-SSN – Case Study

Understand the placement of **MonF** in the **sequence–function space of PTases**





How to Use EFI-SSN – Case Study

BlueBio Future

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

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

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

The database is selected using the BLAST Retrieval Options.

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

Query Sequence:

```
>MonF
MLTCNNILNSEAKLHAIGTHKNAFEIESLYPLDIFEQLVAQTGECGLEFSCKIERERL
YPARFNLSFGRD
RKFADSFKQILSFFHQVAGRVGTTINYELLQEFIGDNFDWNQVEGLQTGVDLRPE
VTNSRLKLLFRIKDY
```

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

BLAST Retrieval Options

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

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

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

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

SSN Edge Calculation Option

Protein Family Addition Options

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

Familes: IPR031037

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

Family	Family Name	Full Size	UniRef90 Size	UniRef50 Size
IPR031037	Preny_LynF_TruF	154	91	40
	Total:	154	91	40
	Total Computed:	**154**		

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.

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 UniProtKB, so the selected sequences are assumed to be a representative sampling of the family. This allows reduction of the size of the SSN. Sequences in the family with SwissProt annotations will always be included; this may result in the size of the resulting data set being slightly larger than the fraction specified.

Job name: *MonF* (required)

E-mail address: *raquelcastelobranco12@gmail.com*

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

Submit Analysis

How to Use EFI-SSN – Case Study

BlueBio Future

DATASET COMPLETED

Submission Name: **MonF**

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

Dataset Summary | Taxonomy Sunburst | Dataset Analysis | SSN Finalization

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

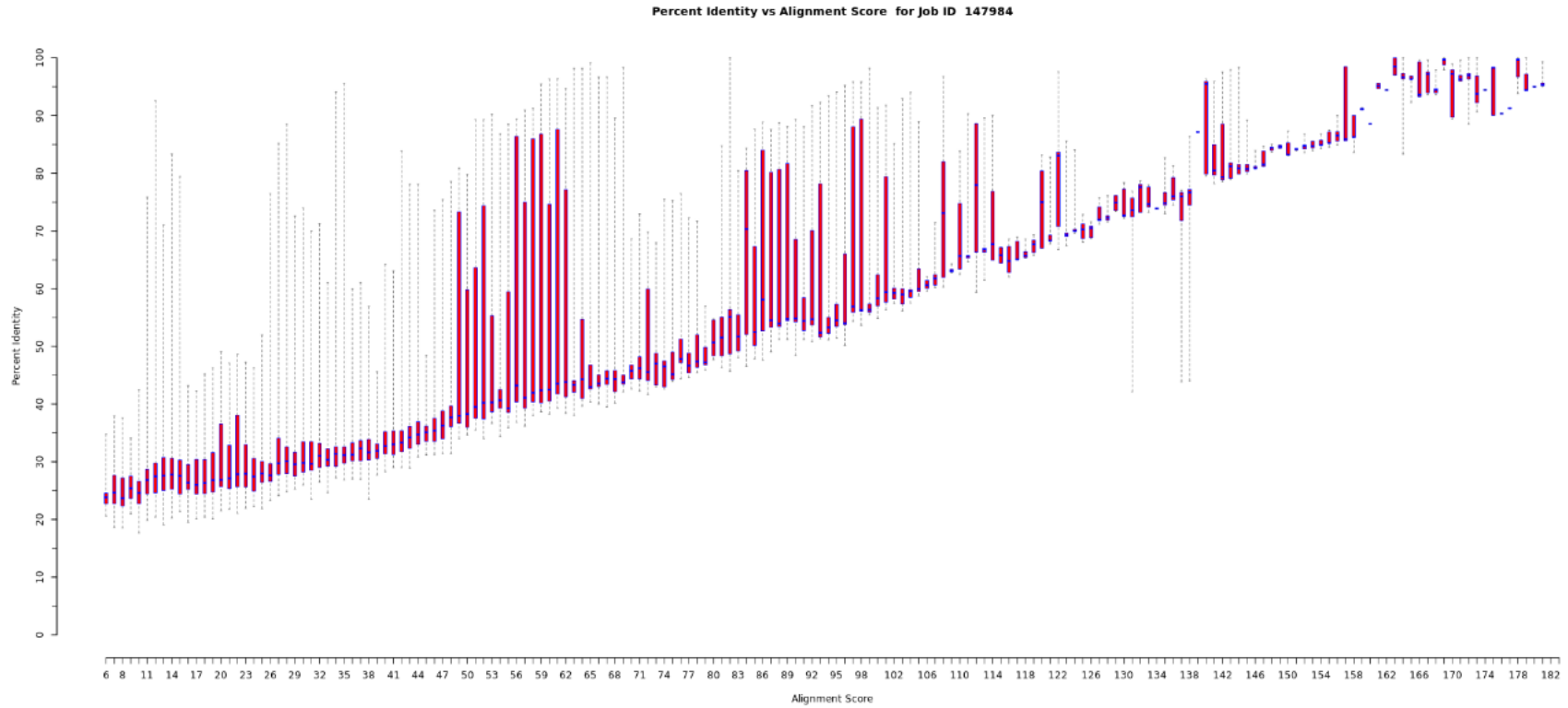
Job Number	147984
Time Started -- Finished	3/9 01:33 PM -- 3/9 02:04 PM
Database Version	UniProt: 2024-05 / InterPro: 102
Input Option	Sequence BLAST (Option A)
Job Name	MonF
E-Value for UniProt BLAST Retrieval	5
E-Value for SSN Edge Calculation	5
Sequence Submitted for BLAST	**View Sequence**
BLAST Database	UniProt
Maximum Number of Retrieved Sequences	1,000
Actual Number of Retrieved Sequences	141
Exclude Fragments	No
Total Number of Sequences in Dataset	142 (includes sequence submitted for BLAST)
Total Number of Edges	9,537
Number of Unique Sequences	129
Convergence Ratio ?	0.953

Download Information

How to Use EFI-SSN – Case Study

BlueBio Future

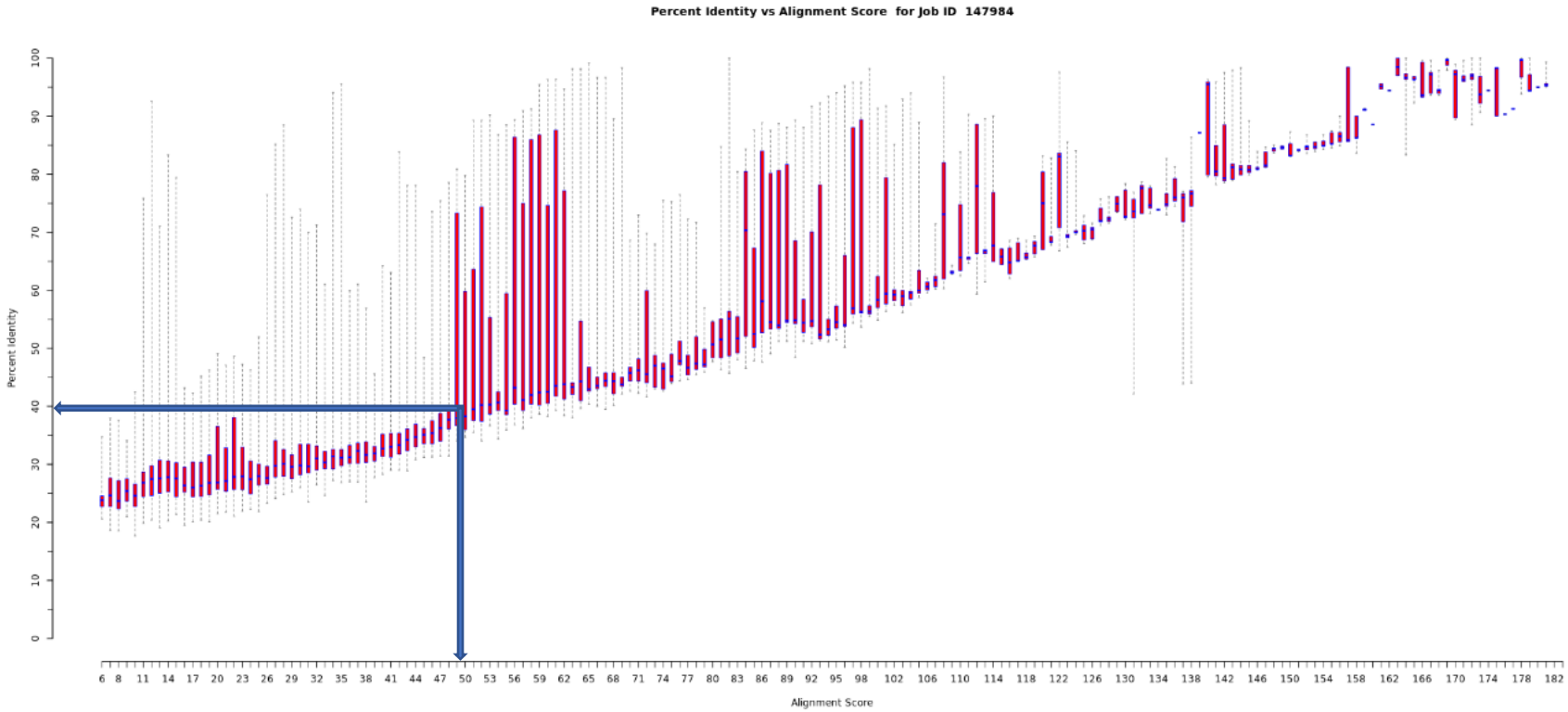
Percent Identity vs Alignment Score Box Plot (Third Step for Alignment Score Threshold Selection)



How to Use EFI-SSN – Case Study

BlueBio Future

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



How to Use EFI-SSN – Case Study

BlueBio Future

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

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

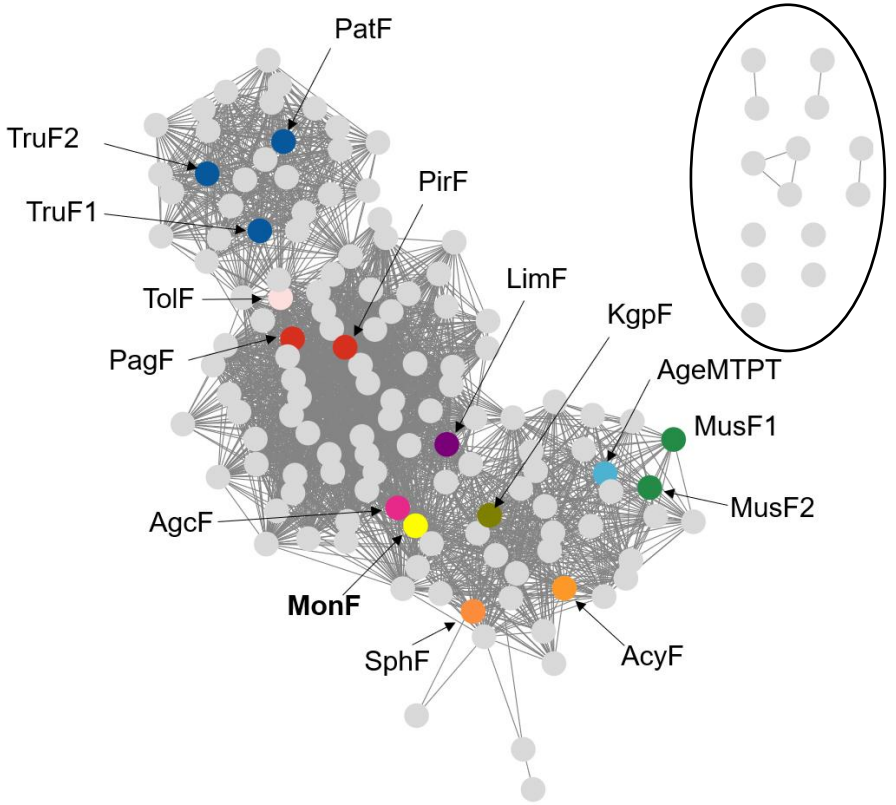
Alignment Score Threshold: 50

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.

Cytoscape

https://cytoscape.org/

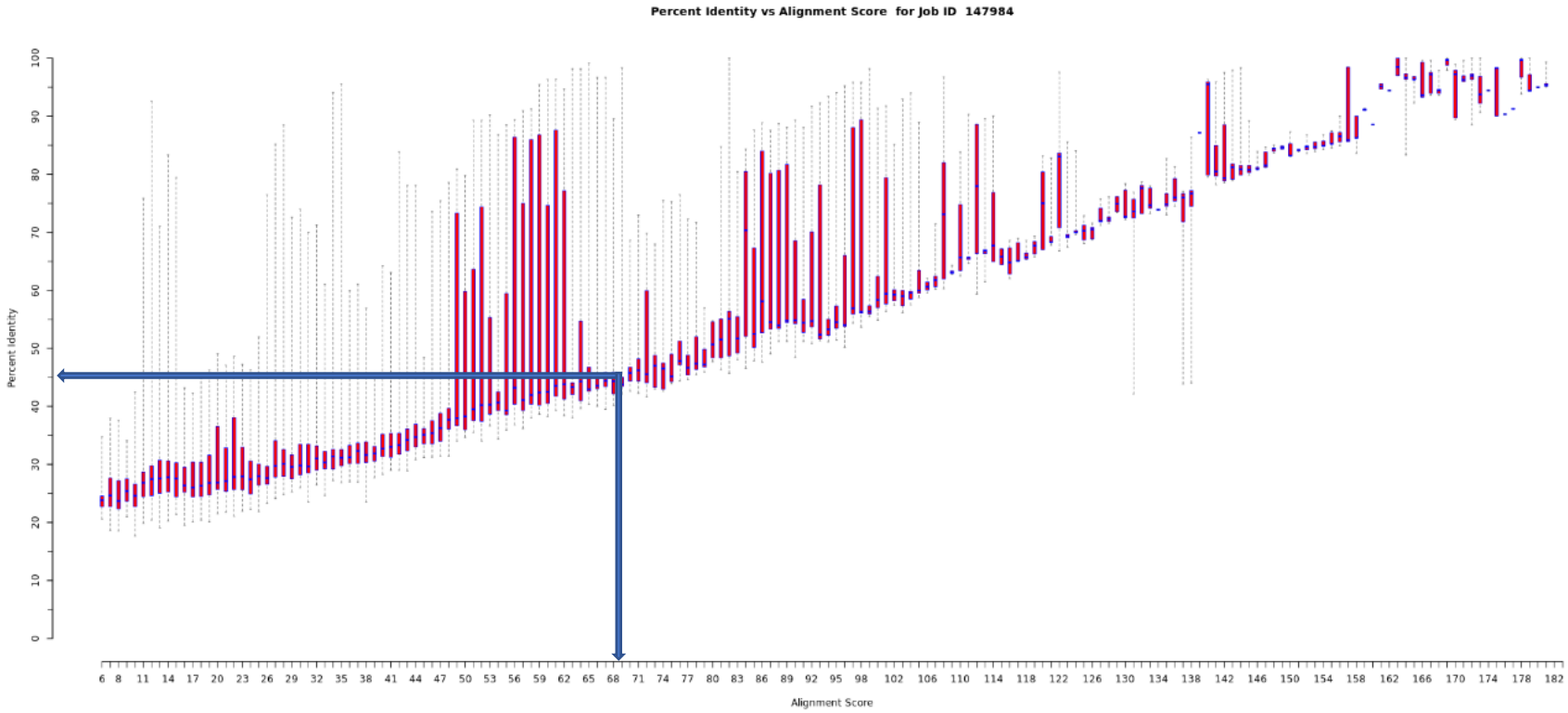
Non-cyanobacterial hits



How to Use EFI-SSN – Case Study

BlueBio Future

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



How to Use EFI-SSN – Case Study

BlueBio Future

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

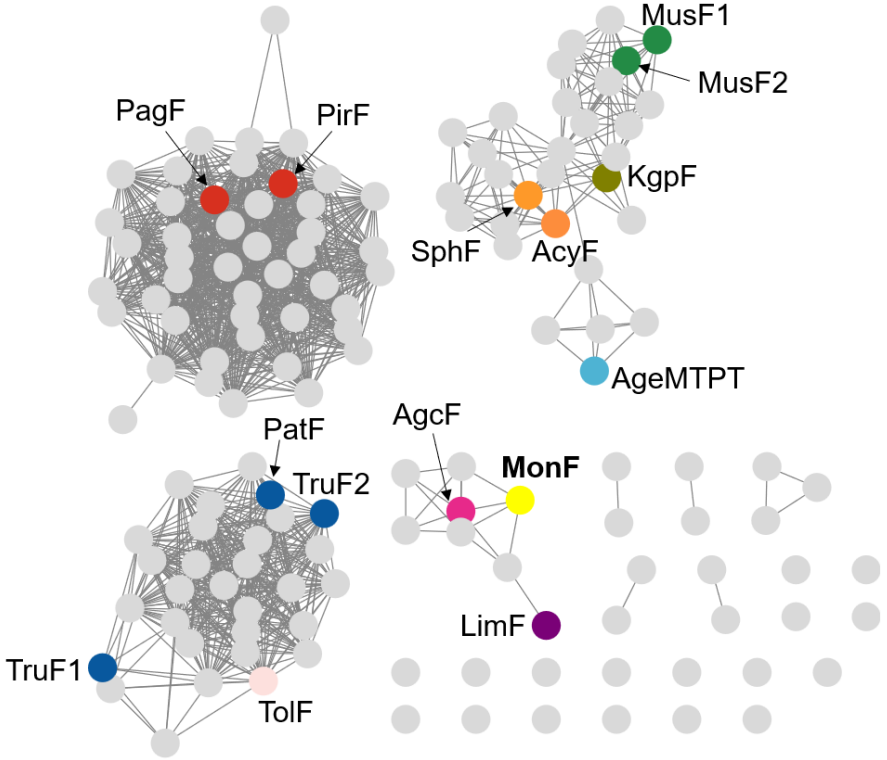
This tab is used to specify the minimum "Alignment Score Threshold" (that is a measure of the minimum sequence similarity threshold) for drawing the edges that connect the proteins (nodes) in the SSN.

Alignment Score Threshold: 68 ?

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.

Cytoscape

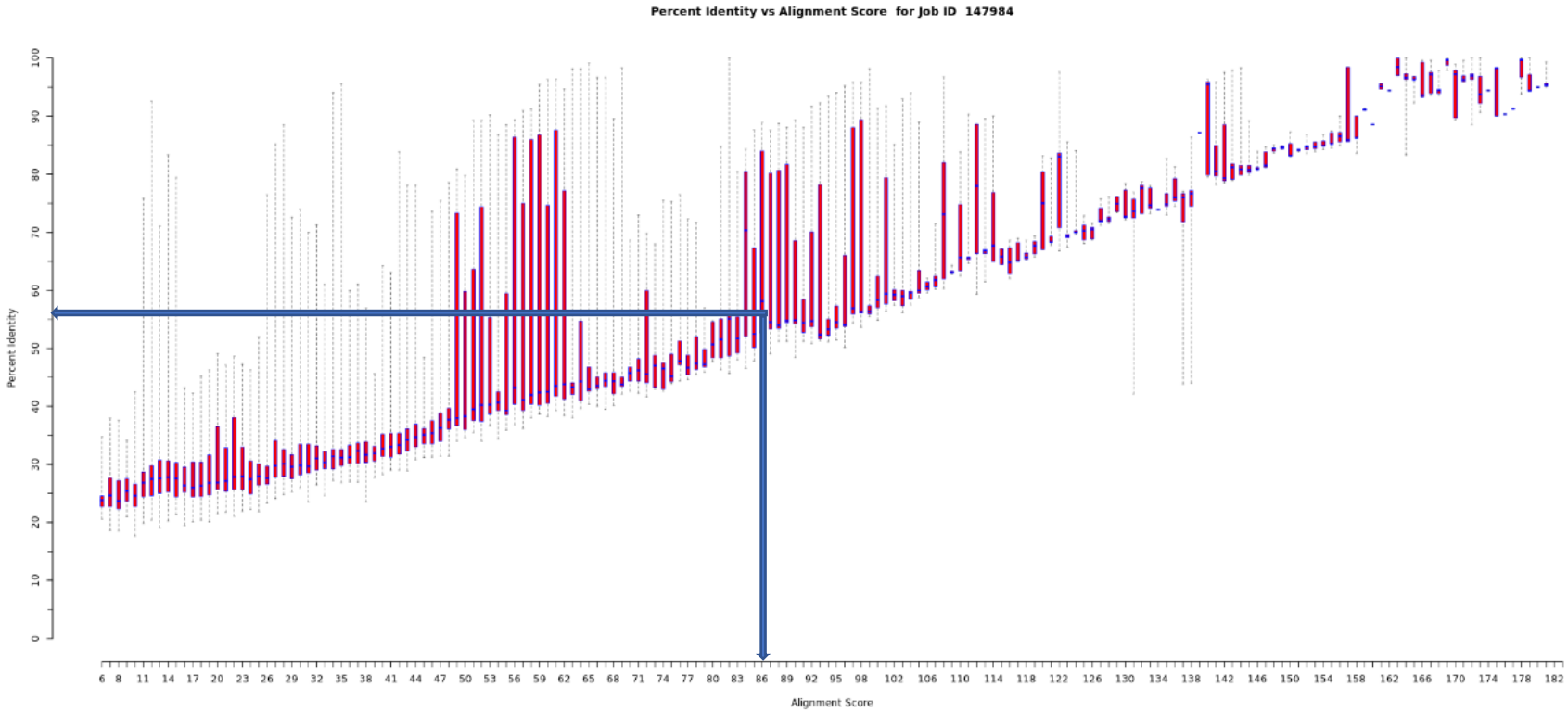
https://cytoscape.org/



How to Use EFI-SSN – Case Study

BlueBio Future

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



How to Use EFI-SSN – Case Study

BlueBio Future

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

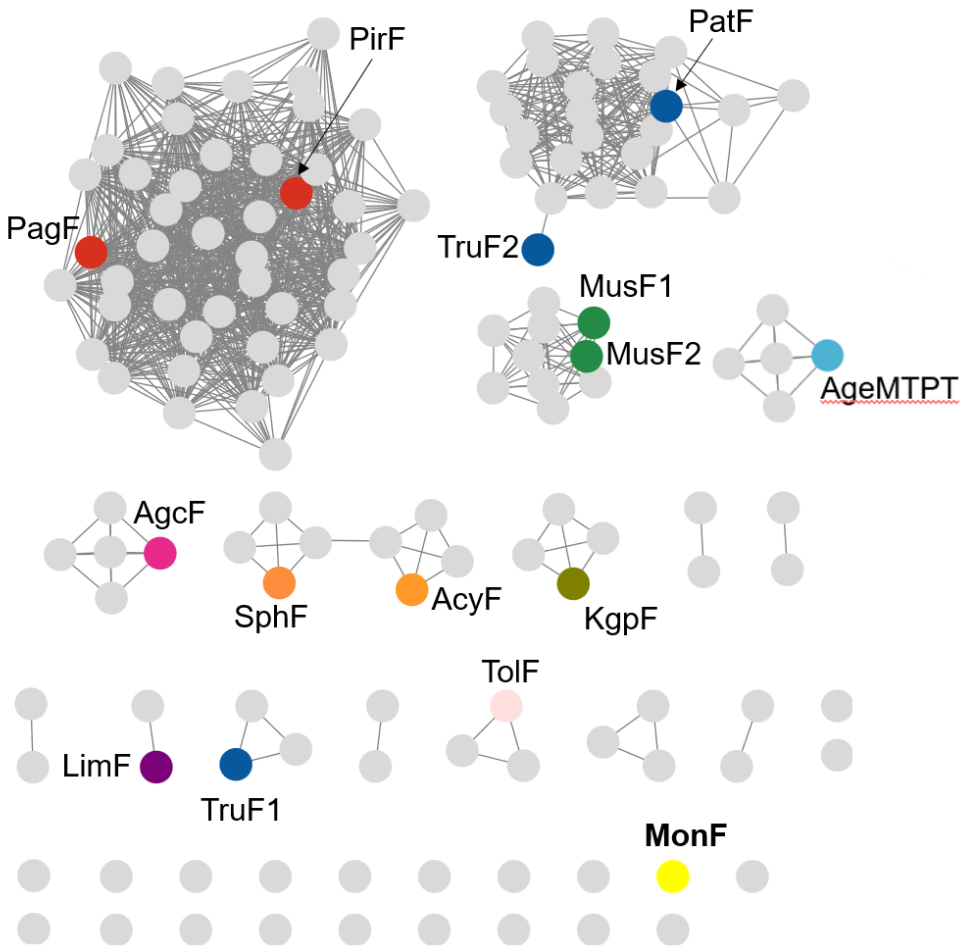
This tab is used to specify the minimum "Alignment Score Threshold" (that is a measure of the minimum sequence similarity threshold) for drawing the edges that connect the proteins (nodes) in the SSN.

Alignment Score Threshold: 86 ?

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.

Cytoscape

https://cytoscape.org/

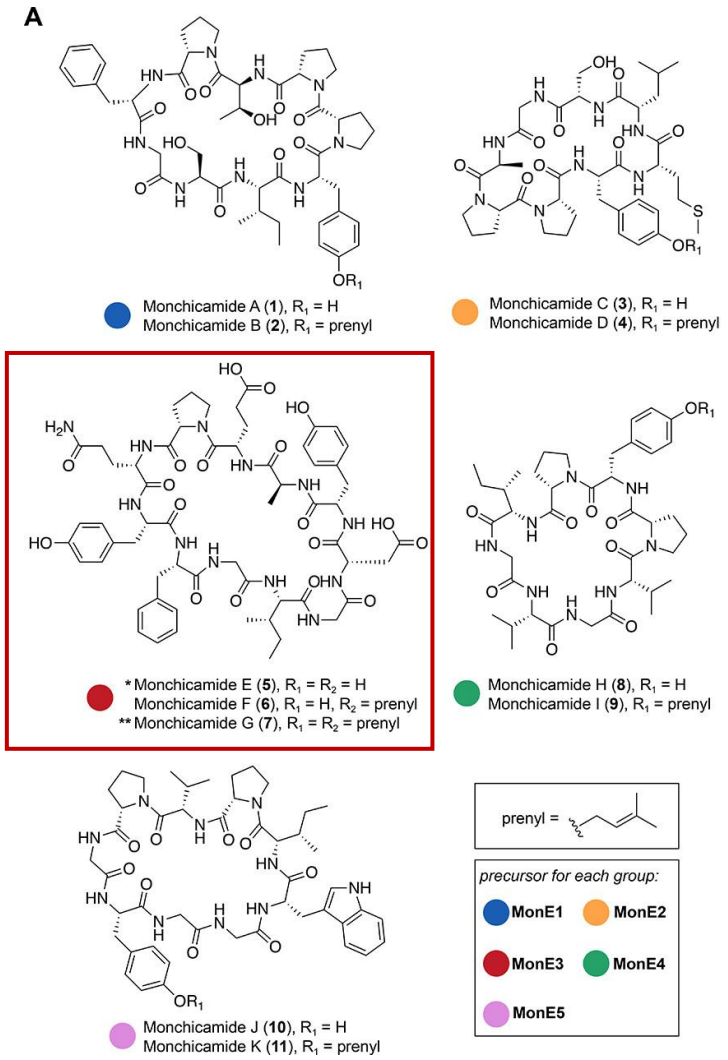


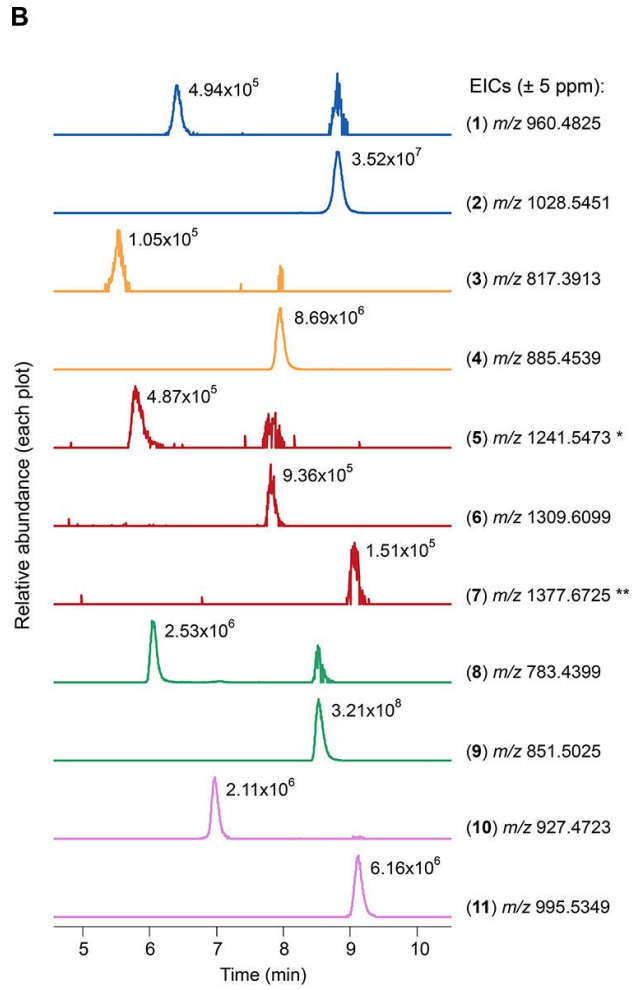
How to Use EFI-SSN

MonF is a Tyr-*O*-Ptase (forward orientation)

Monchicamide G presents appears to be the first cyanobactin reported with **prenyl group** in **two Tyr** residues catalyzed by **a single Ptase**.

Biochemical characterization of MonF is required to confirm if catalyzes both prenylations and to investigate the substrate specificity.





Conclusions

Advantages of EFI-SSN

- ☑ **Scalability**
  Processes large datasets with thousands of sequences.

- ☑ **Visualization**
  Provides an interactive way to explore enzyme families.

- ☑ **User-Friendly**
  Web-based tool with an easy-to-use interface.

- ☑ **Integration with Other Tools**
  Supports cross-platform compatibility with other computational enzyme discovery methods.

EFI SSN is a powerful tool for enzyme discovery, enabling functional prediction and classification of protein families.

Acknowledgments

Thank you!

rcastelobranco@ciimar.up.pt

Funded by the European Union
European Research Council
Fundação para a Ciência e a Tecnologia
PORTUGAL 2020
NORTE2020
REPÚBLICA PORTUGUESA
ciimar Centro Interdisciplinar de Investigação Marinha e Ambiental
cyano natural products team