Large Scale Enzyme Function Discovery: Sequence Similarity Networks for the "Protein Universe"

Boris Sadkhin

University of Illinois, Urbana-Champaign

Blue Waters Symposium

May 2015

Overview

- The Protein Sequence Database Problem
- Sequence Similarity Networks (SSNs)
- EFI-EST (Enzyme Similarity Tool)
- EST-Precompute

Personnel involved in this project

Carl R. Woese Institute for Genomic Biology (IGB) at University of Illinois, Urbana-Champaign

John A. Gerlt, PI
Victor Jongeneel, CoPI
Daniel Davidson
David Slater

External Collaborators

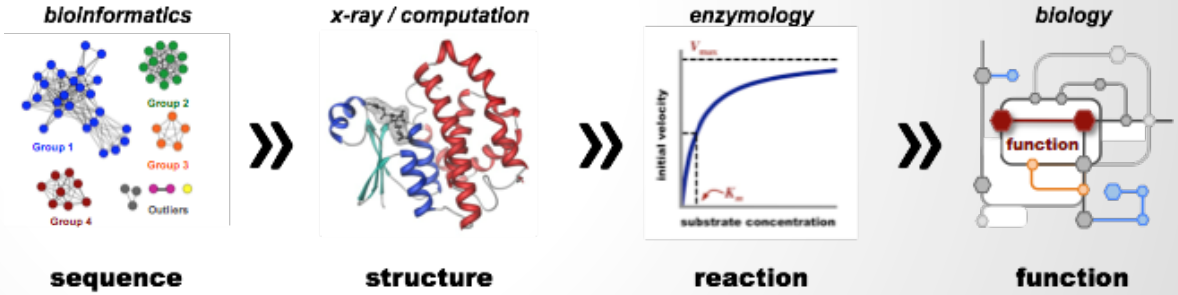
Alex Bateman, EMBL-EBI
Matthew Jacobson, UCSF

The Enzyme Function Initiative (EFI)

- The Enzyme Function Initiative, an NIH/NIGMS−supported Large−Scale Collaborative Project (EFI; U54GM093342; http://enzymefunction.org/)

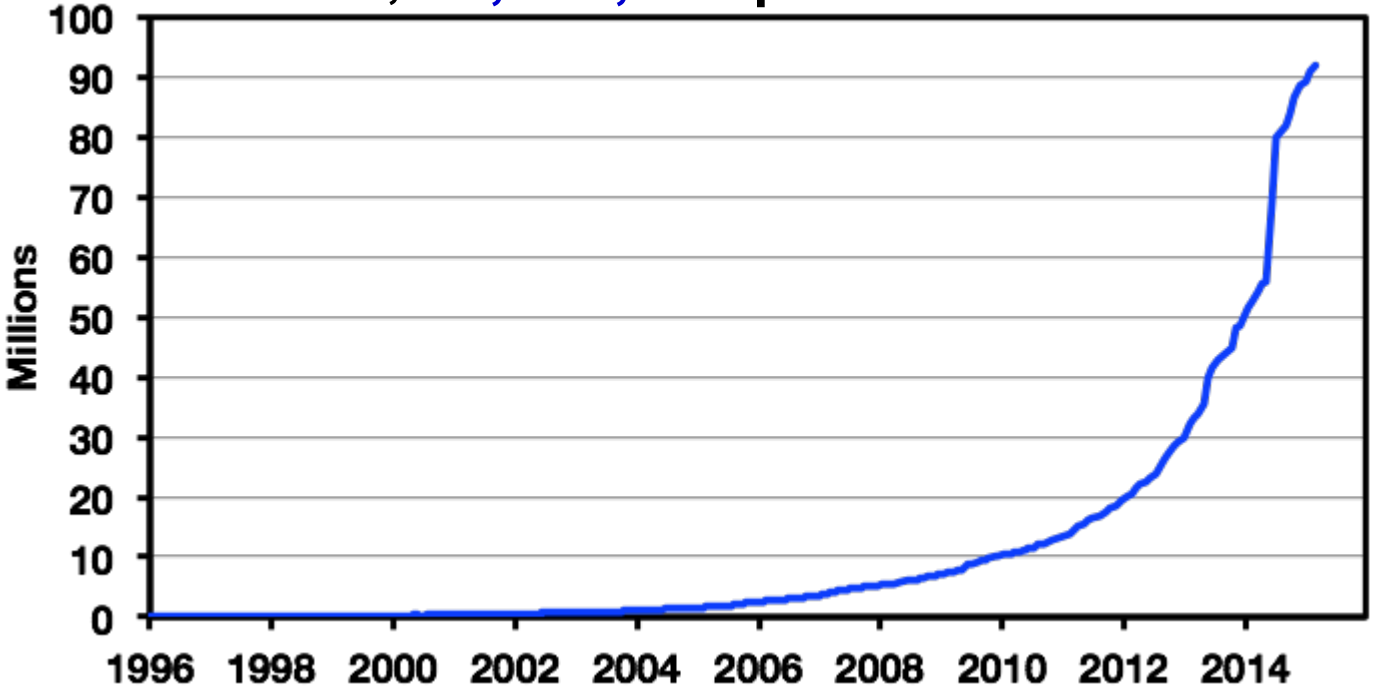
What do we do?

- ***collaborate***
- ***create***
- ***disseminate***



An explosion of protein sequences!

As of March 2015, 92,124,243 proteins had been identified.



The Problem

- The number of protein sequences is ***exploding!***
- 50% of our protein databases are ***misannotated!***
- There are many proteins and enzymes to ***discover!***





The Solution

A Sequence Similarity Network Database

EFI EST | PRECOMPUTE

NCSA

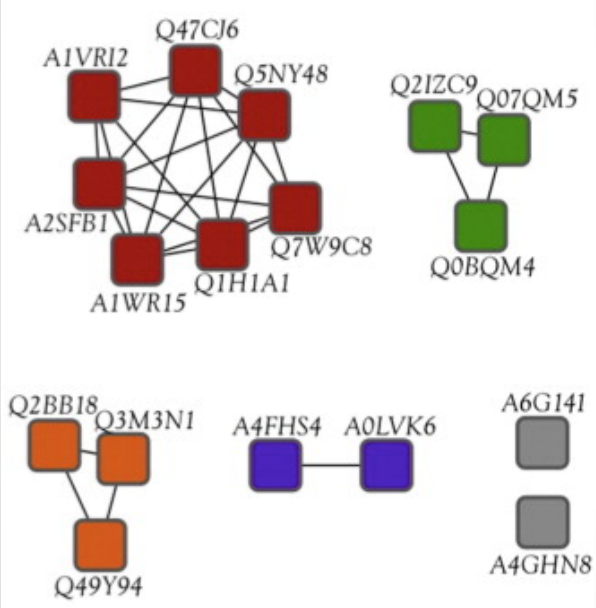
Bridging the Gap : Biologists and Big Data

NCSA + EFI ENZYME FUNCTION INITIATIVE

Generating the database on BW

	Biocluster @ IGB	Blue Waters @ NCSA
# of Nodes	20 EFI Nodes @24 cpu 20 Shared Nodes @24 cpu	> 22,000 Nodes @ 32 cpu
Storage (100TB)	600 TB for entire cluster	500 TB for just our project
>90 million sequences **=4,243,438,028,099,403 comparisons**	8 months	< 2 weeks
Node hours?	• 200,000 node hours • 6,400,000 cpu hours	

What is a Sequence Similarity Network?



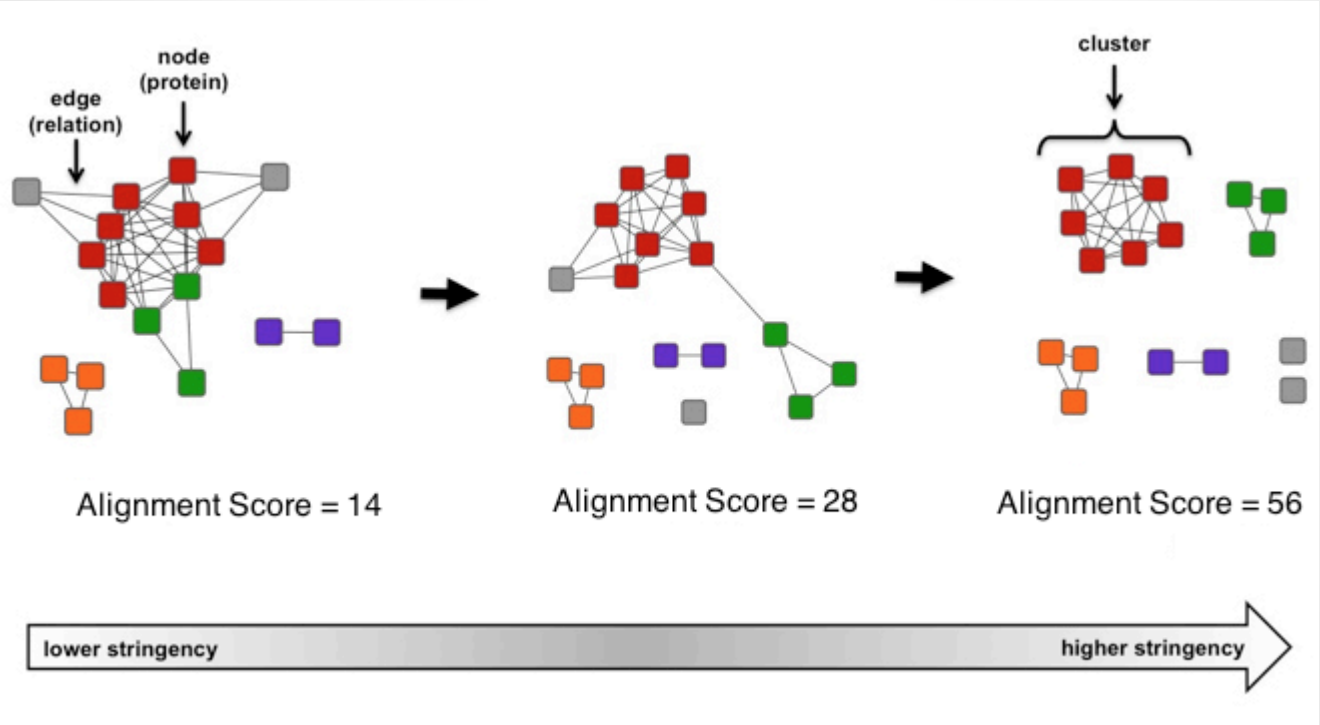
node (circle) = protein sequence

edge (line) = alignment score

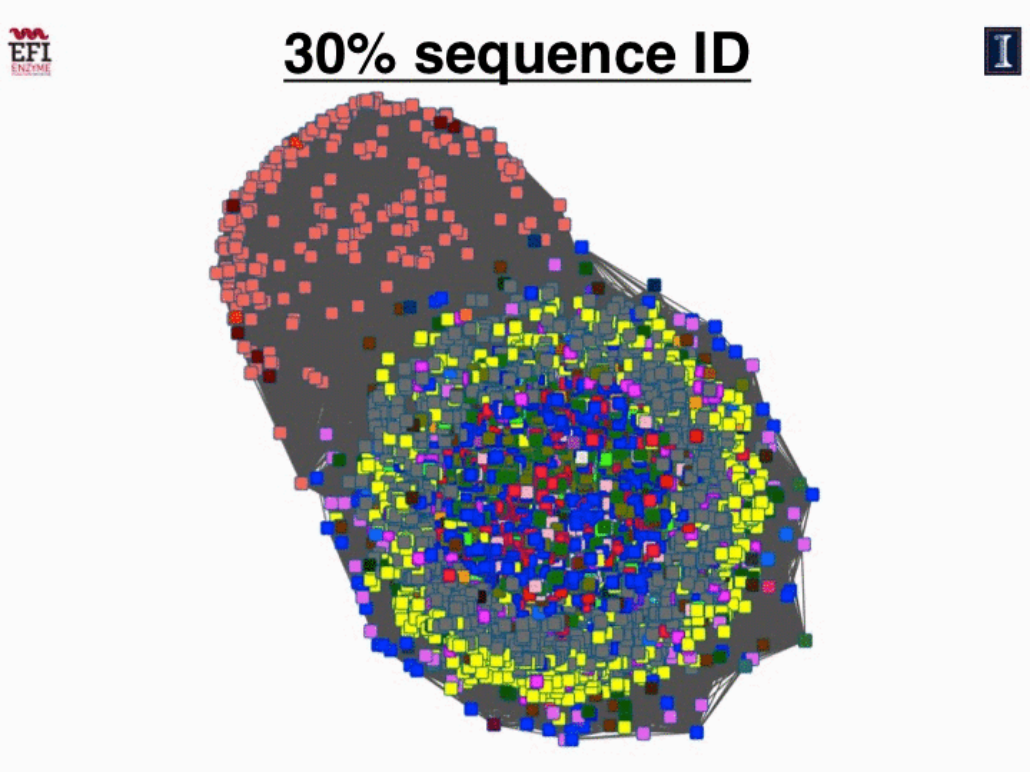
$$-\log_{10}\left[2^{-\text{bitscore}} \cdot (\text{query length} \cdot \text{subject length})\right]$$

Alignment Score

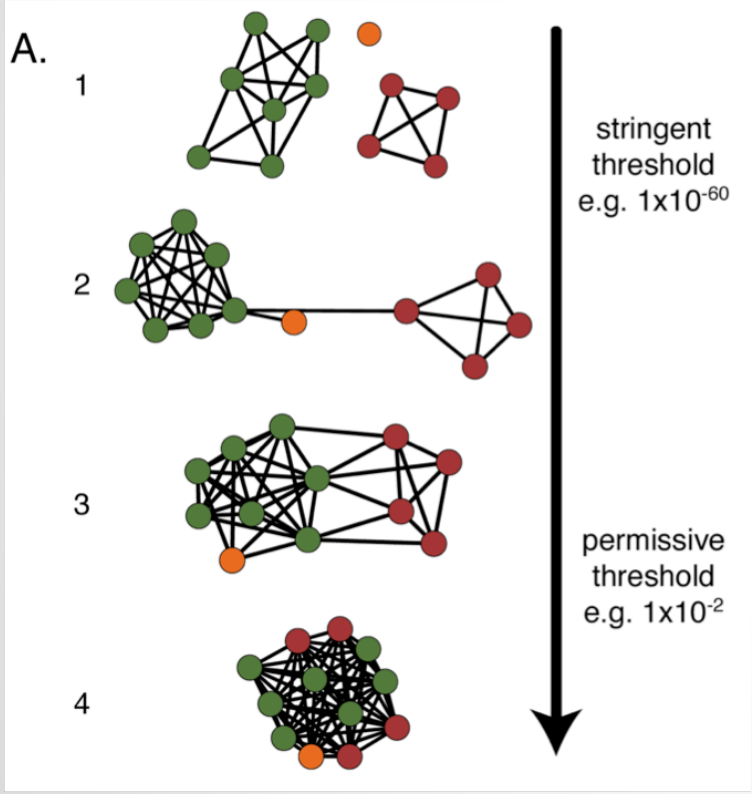
Using Sequence Similarity Networks

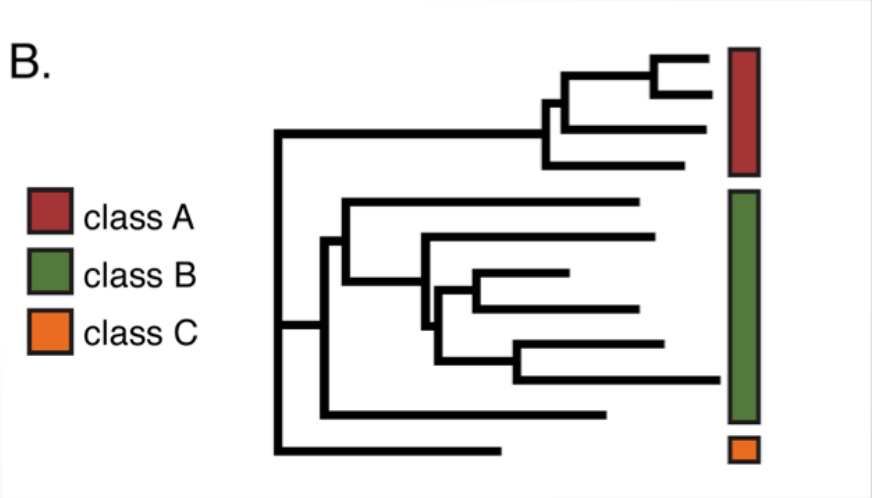


Using Sequence Similarity Networks



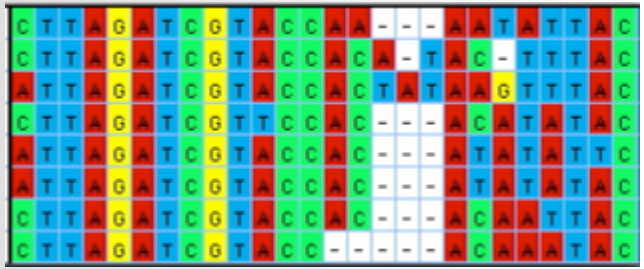
SSNS- Computationally Faster, Qualitatively Similar



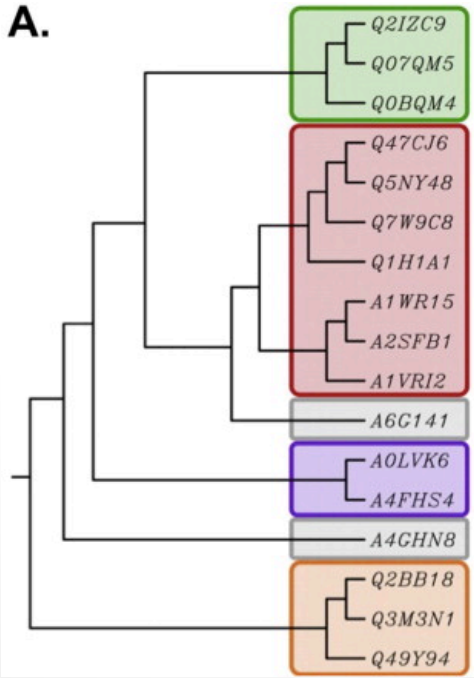


Analyzing Groups of Proteins

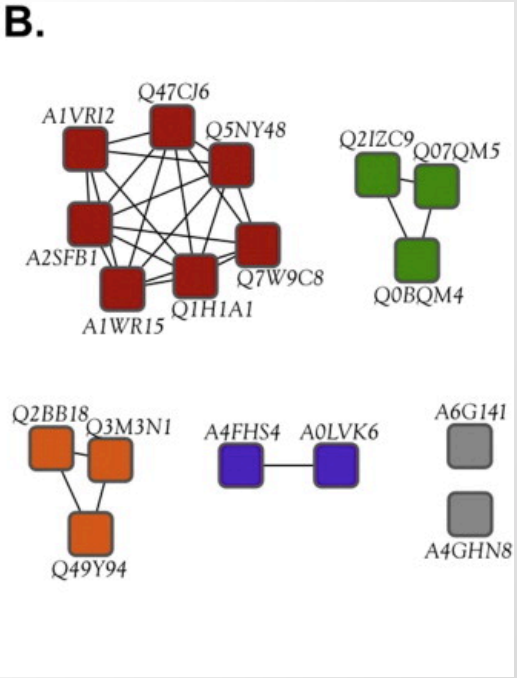
Multiple Sequence Alignment



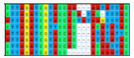
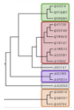
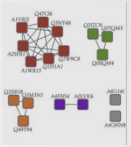
Phylogenetic Trees and Dendrograms



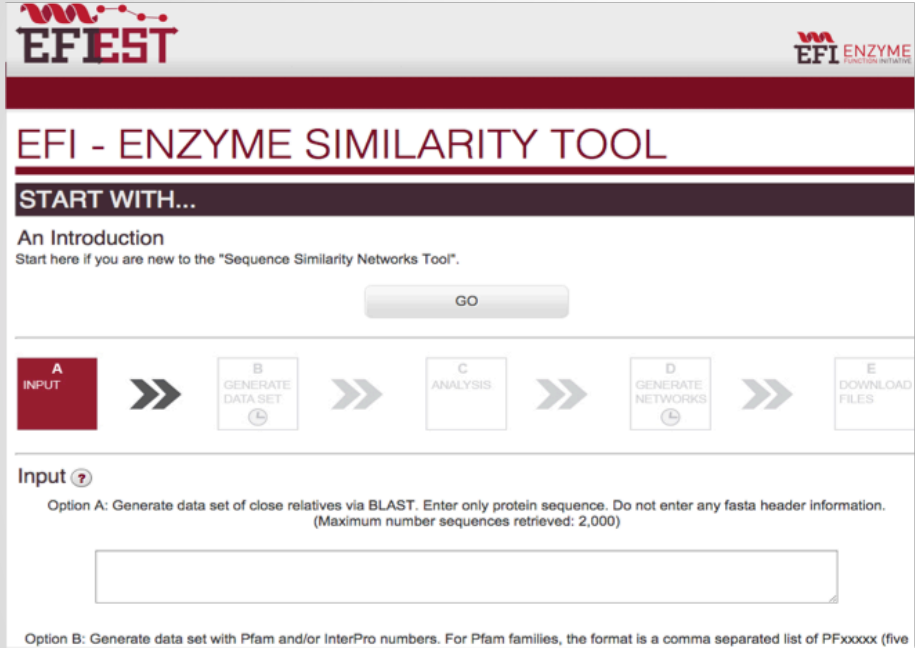
Sequence Similarity Networks

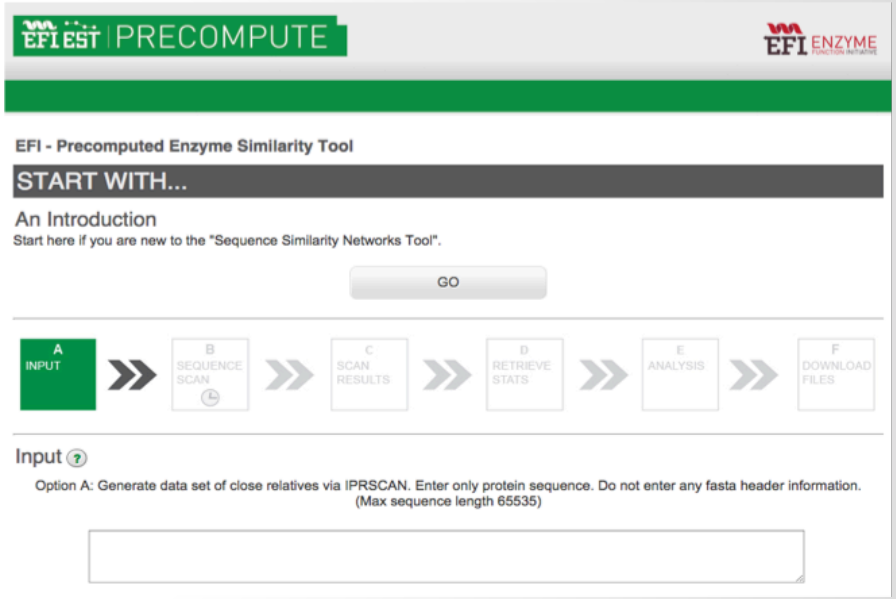


Pros and Cons

	Multiple Sequence Alignment (MSA) 	Phylogenetic Trees 	Sequence Similarity Networks (SSNs) 
Visualization of Small Datasets	Good ✓	Good ✓	Good ✓
Visualization of Large Datasets	Bad	Not so good	Good ✓✓✓
Informative	Small Datasets ✓ Large Datasets X	Small Datasets ✓ Large Datasets X	Small Datasets ✓ Large Datasets ✓
Computational Cost	Expensive	Requires Sensitive MSA	Pairwise Sequence Alignment BLAST heuristics
Displays Annotations?	No	Sometimes	***26*** (eg...crosslinks)

Our SSN Tools





EFIEST

EFI ENZYME FUNCTION INITIATIVE

EFI - ENZYME SIMILARITY TOOL

START WITH...

An Introduction

Start here if you are new to the "Sequence Similarity Networks Tool".

GO

A INPUT » B GENERATE DATA SET » C ANALYSIS » D GENERATE NETWORKS » E DOWNLOAD FILES

Input

Option A: Generate data set of close relatives via BLAST. Enter only protein sequence. Do not enter any fasta header information. (Maximum number sequences retrieved: 5,000).

To convert your blast search into an InterPro number, please go to **http://www.ebi.ac.uk/interpro/**

Option B: Generate data set with Pfam and/or InterPro numbers. For Pfam families, the format is a comma separated list of PFxxxxx (five digits); for InterPro families, the format is IPRxxxxxx (six digits). (Maximum number sequences retrieved: 100,000)

Enter your email address

Used for data retrieval only

GO

efi.igb.illinois.edu/efi-est/

- Enzyme Similarity Tool

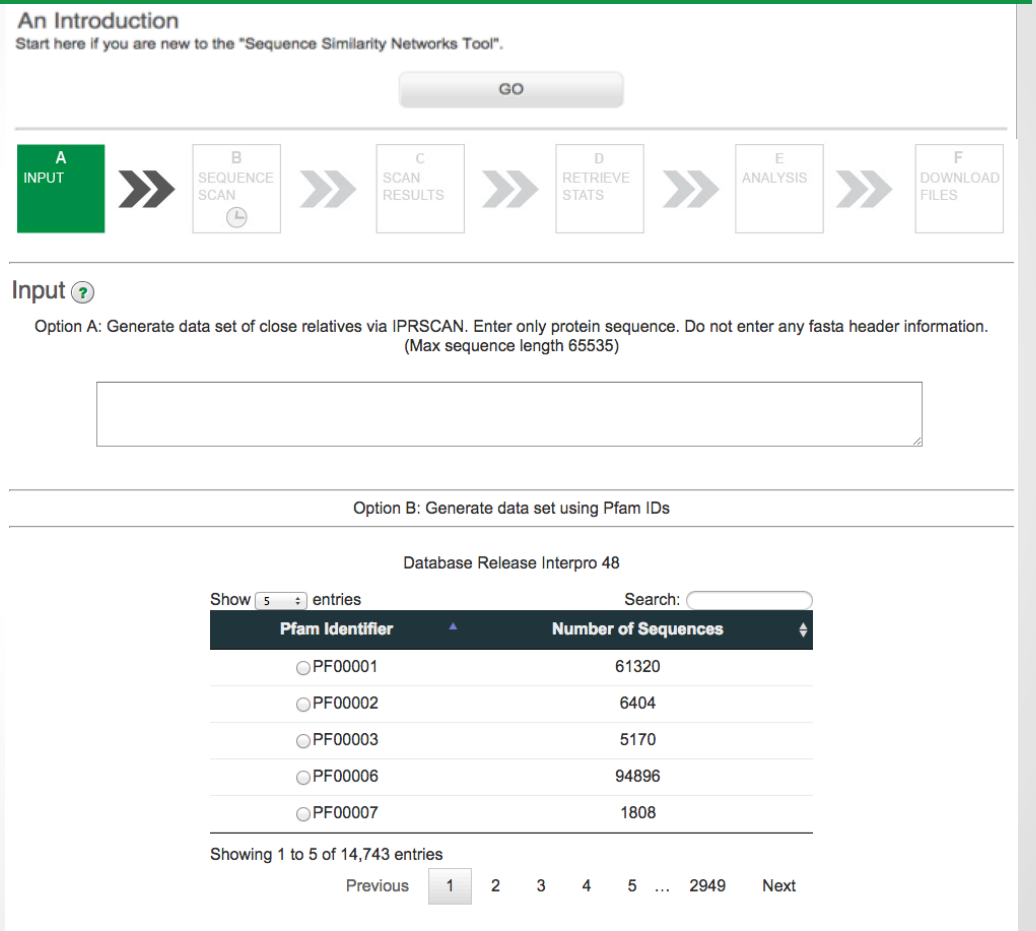
Caveats:

- 100,000 sequence threshold for predefined families
- Takes time, networks need to be generated and regenerated for filtering

EFI EST | PRECOMPUTE

- Gene3D
- PFAM Clans
- Interpro Families
- More?

efi.igb.illinois.edu/est-precompute



EFI EST | PRECOMPUTE

DOWNLOAD NETWORK FILES

Full Network

Each node in the network is a single protein from the data set. Large files (>500MB) may not open.

Filename		# Nodes	# Edges	XGMML Size	Zipped
Download PF00003-full.xgmml.gz	Full	5,170	7,438,875	1.87G	181.94M

Representative Node Networks

Each node in the network represents a collection of proteins grouped according to percent identity.

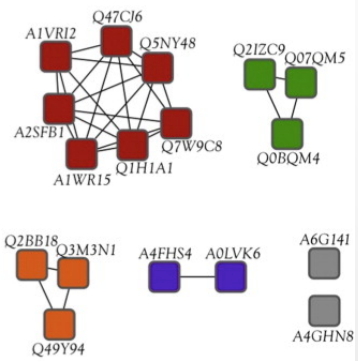
Filename	% ID	# Nodes	# Edges	XGMML Size	Zipped
Download PF00003-40.xgmml.gz	40	503	43,367	14.40M	1.61M
Download PF00003-45.xgmml.gz	45	639	73,556	22.57M	2.57M
Download PF00003-50.xgmml.gz	50	792	118,051	34.41M	3.97M
Download PF00003-55.xgmml.gz	55	938	173,033	48.88M	5.67M
Download PF00003-60.xgmml.gz	60	1,131	264,200	72.75M	8.47M
Download PF00003-65.xgmml.gz	65	1,330	378,467	102.56M	11.95M
Download PF00003-70.xgmml.gz	70	1,543	535,430	143.32M	16.70M
Download PF00003-75.xgmml.gz	75	1,819	782,646	207.37M	24.12M
Download PF00003-80.xgmml.gz	80	2,144	1,123,292	295.33M	34.28M
Download PF00003-85.xgmml.gz	85	2,519	1,609,094	420.54M	48.65M
Download PF00003-90.xgmml.gz	90	2,961	2,305,740	599.81M	69.08M
Download PF00003-95.xgmml.gz	95	3,576	3,468,600	898.64M	102.63M
Download PF00003-98.xgmml.gz	98	4,112	4,649,669	1.17G	136.21M
Download PF00003-100.xgmml.gz	100	4,924	6,785,403	1.71G	170.09M

Full SSNs

- each node = 1 sequence

Representative SSNs

- each node > 1 sequence



EST & EST-Precompute use

- widely used database of conserved protein families that are based on a seed alignment of representative sequences that are used to generate a profile hidden Markov model (HMM)

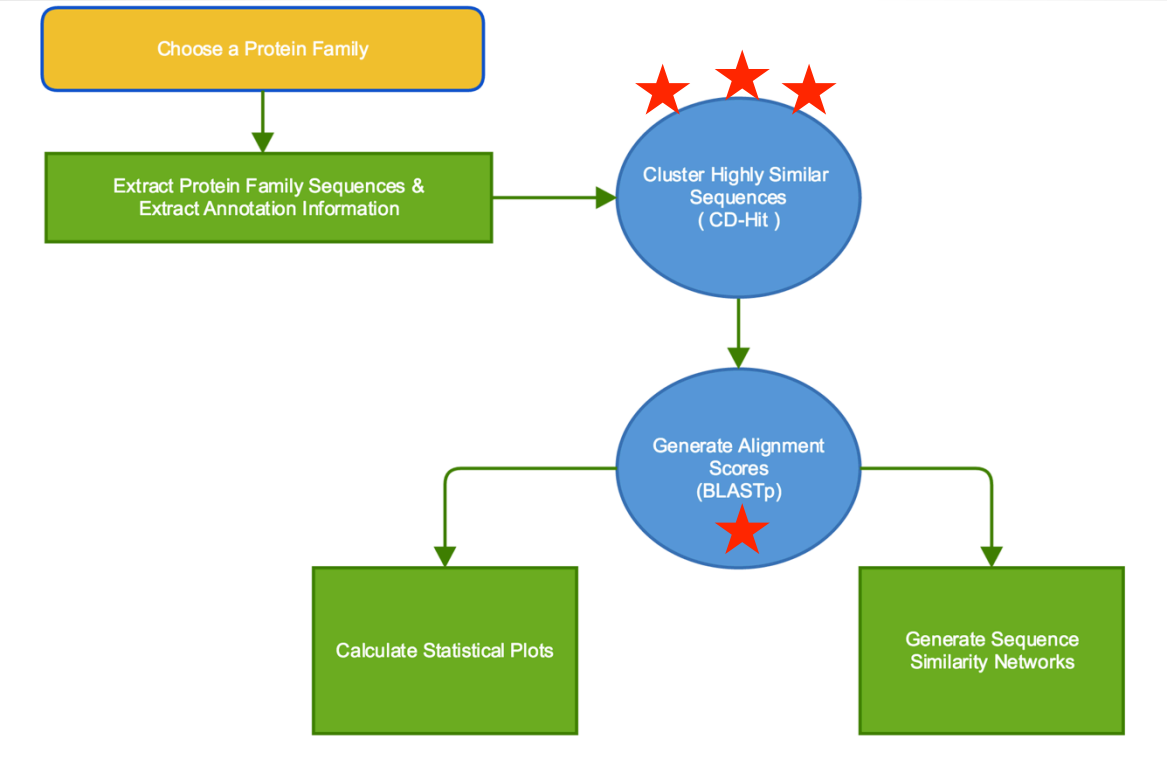
- **14,831 defined families in Pfam**

http://pfam.xfam.org/

Challenges:

- The “doubling time” of the UniProt database (http://www.uniprot.org/), is ~ 18 months
- Adapting the workflow and algorithms for increasingly large sequence datasets
- Dealing with major changes in the databases from which we get our data

Our Workflow



Accomplishments

- Dealing with the ‘explosion’ of protein sequences
- Algorithms
- Generated > 14,000 Pfams
- Production Pipeline

Blue Waters Team Contributions

The Blue Waters Team has been helpful in dealing with our issues

- Live chat support
- Supplying job stats, optimizing our workflow, fixing software installations, you name it
- scheduler.x - the single threaded job scheduler

Thank You!

Questions?

References

Sequence Similarity Networks in the SFLD	
EFI EST	http://www.sciencedirect.com/science/article/pii/S1570963915001120
Pfam	R.D. Finn, A. Bateman, J. Clements, P. Coggill, R.Y. Eberhardt, S.R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E.L. Sonnhammer, J. Tate, and M. Punta, Pfam: the protein families database. Nucleic Acids Res 2014, 42, D222-30. PMCID: PMC3965110
Uniprot	C. UniProt UniProt: a hub for protein information Nucleic Acids Res, 43 (2015), pp. D204–D212
Collaborator Patsy Babbitt	http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2781113/ [4]
PMC	http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1892569/ [5]

EFI's "funnel": strategy for functional assignment

Target Selection
Bioinformatics Core

Protein production and X-ray
Protein / Structure Cores

Modeling / Docking
Modeling Core

Enzymology
Bridging Projects

Genetics / Transcriptomics
Metabolism Core

Metabolomics
Metabolism Core