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
Bridging the Gap: Biologists and Big Data

NCSA + EFI

ENZYME FUNCTION INITIATIVE
Generating the database on **BW**

<table>
<thead>
<tr>
<th></th>
<th>Biocluster @ IGB</th>
<th>Blue Waters @ NCSA</th>
</tr>
</thead>
</table>
| **# 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** | 8 months | < 2 weeks |
| **=4,243,438,028,099,403 comparisons** | | |
| **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

30% sequence ID
SSNS- Computationally Faster, Qualitatively Similar

A.  
1.  
2.  
3.  
4.  

stringent threshold e.g. $1 \times 10^{-60}$

B.  

class A

class B

class C
Analyzing Groups of Proteins

Multiple Sequence Alignment

Phylogenetic Trees and Dendrograms

Sequence Similarity Networks
## Pros and Cons

| Visualization of Small Datasets | Good ✓ | Good ✓ | Good ✓ |
| Visualization of Large Datasets | Bad | Not so good | Good ✓ ✓ ✓ |
| Informative | Small Datasets ✓ | Small Datasets ✓ | Small Datasets ✓ |
| | Large Datasets ✗ | Large Datasets ✗ | Large Datasets ✓ |
| Computational Cost | Expensive | Requires Sensitive MSA | Pairwise Sequence Alignment BLAST heuristics |
| Displays Annotations? | No | Sometimes | 26 (eg...crosslinks) |
Our SSN Tools
EFI - ENZYME SIMILARITY TOOL

START WITH...

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

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 PFxxxx (five digits); for InterPro families, the format is IPxxxxx (six digits). (Maximum number sequences retrieved: 100,000)

Enter your email address
Used for data retrieval only

Efili.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
- Gene3D
- PFAM Clans
- Interpro Families
- More?

[efi.igb.illinois.edu/est-precompute]
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

Choose a Protein Family

Extract Protein Family Sequences & Extract Annotation Information

Cluster Highly Similar Sequences (CO-Hit)

Generate Alignment Scores (BLASTp)

Calculate Statistical Plots

Generate Sequence Similarity Networks
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

<table>
<thead>
<tr>
<th>Reference</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence Similarity Networks in the SFLD</td>
<td></td>
</tr>
</tbody>
</table>
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

MGLELFL...
Target

Structure
S → P
Reaction

Physiological Function