Sequence Similarity Networks for the Protein "Universe"

John A. Gerlt

University of Illinois, Urbana-Champaign

Blue Waters Symposium

May 13, 2014

Personnel and Funding

University of Illinois, Urbana-Champaign

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

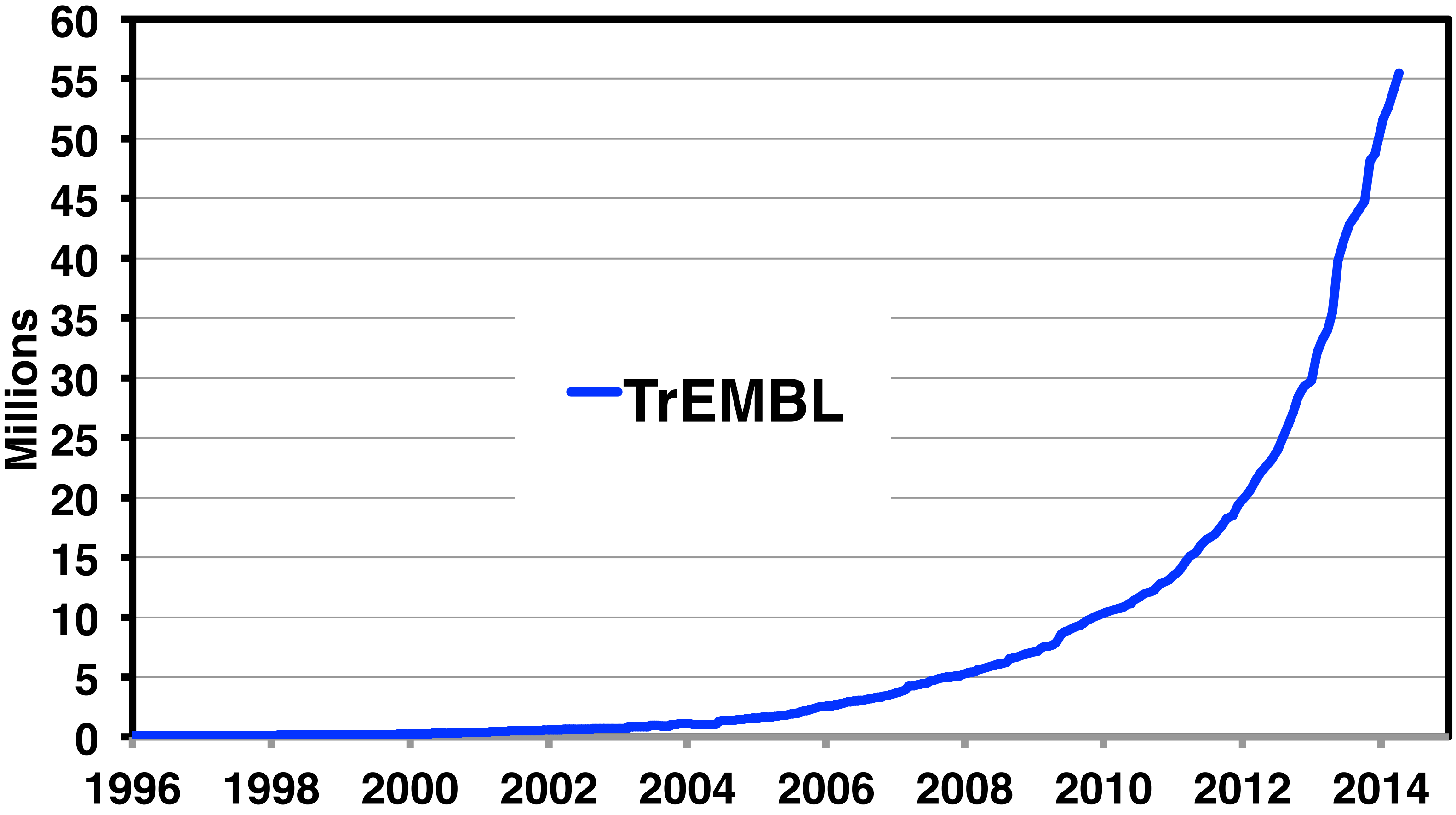
External Collaborators

Alex Bateman, EMBL-EBI
Matthew Jacobson, UCSF

NIH U54GM093342
UIUC Blue Waters Allocation

The number of protein sequences is "exploding" !

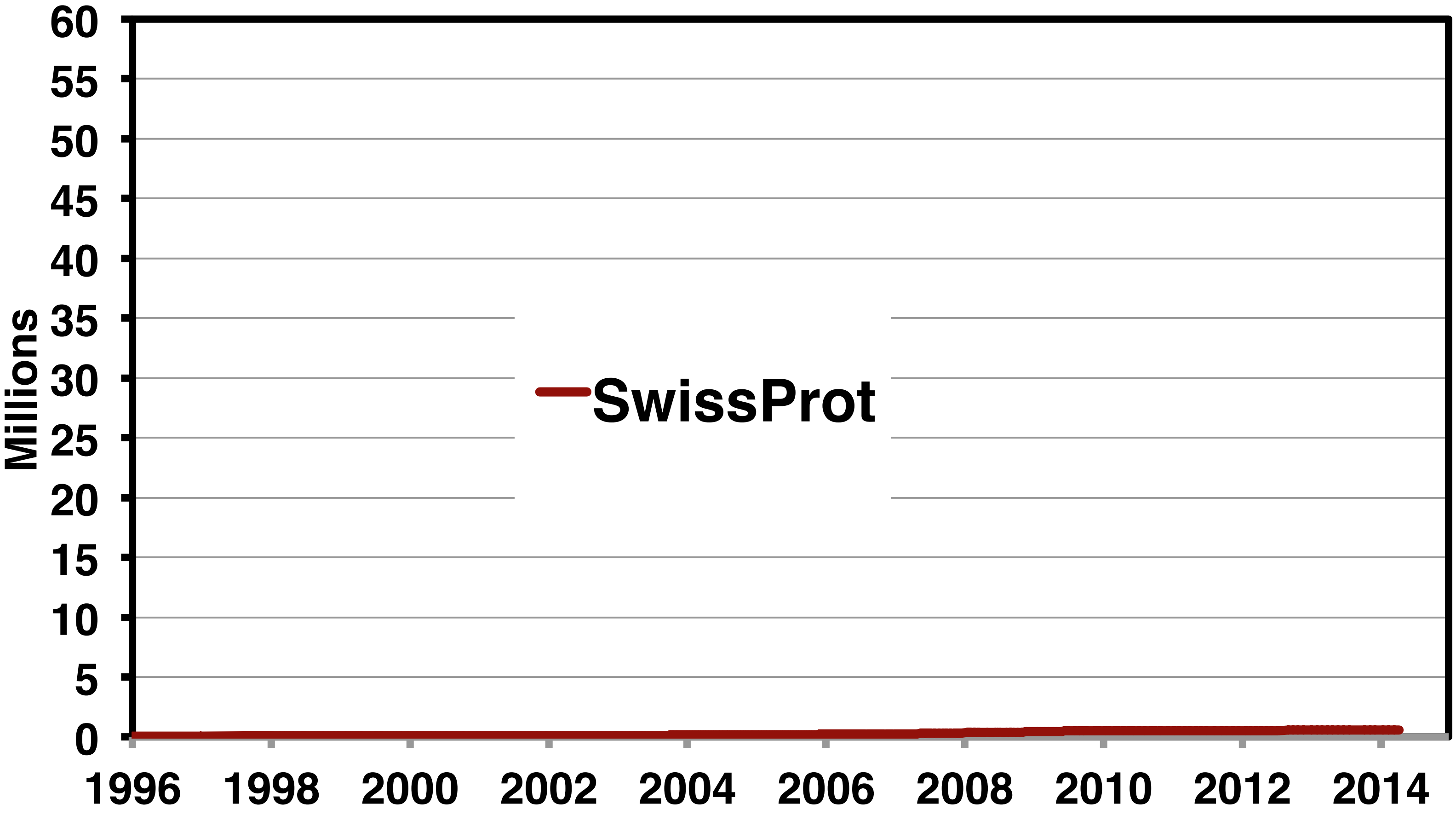
Release 2014_04 of 16-Apr-2014 of UniProtKB/TrEMBL contains 55,503,547 sequence entries.



But the number of curated annotations is lagging !

Release 2014_04 of 16-Apr-2014 of UniProtKB/SwissProt

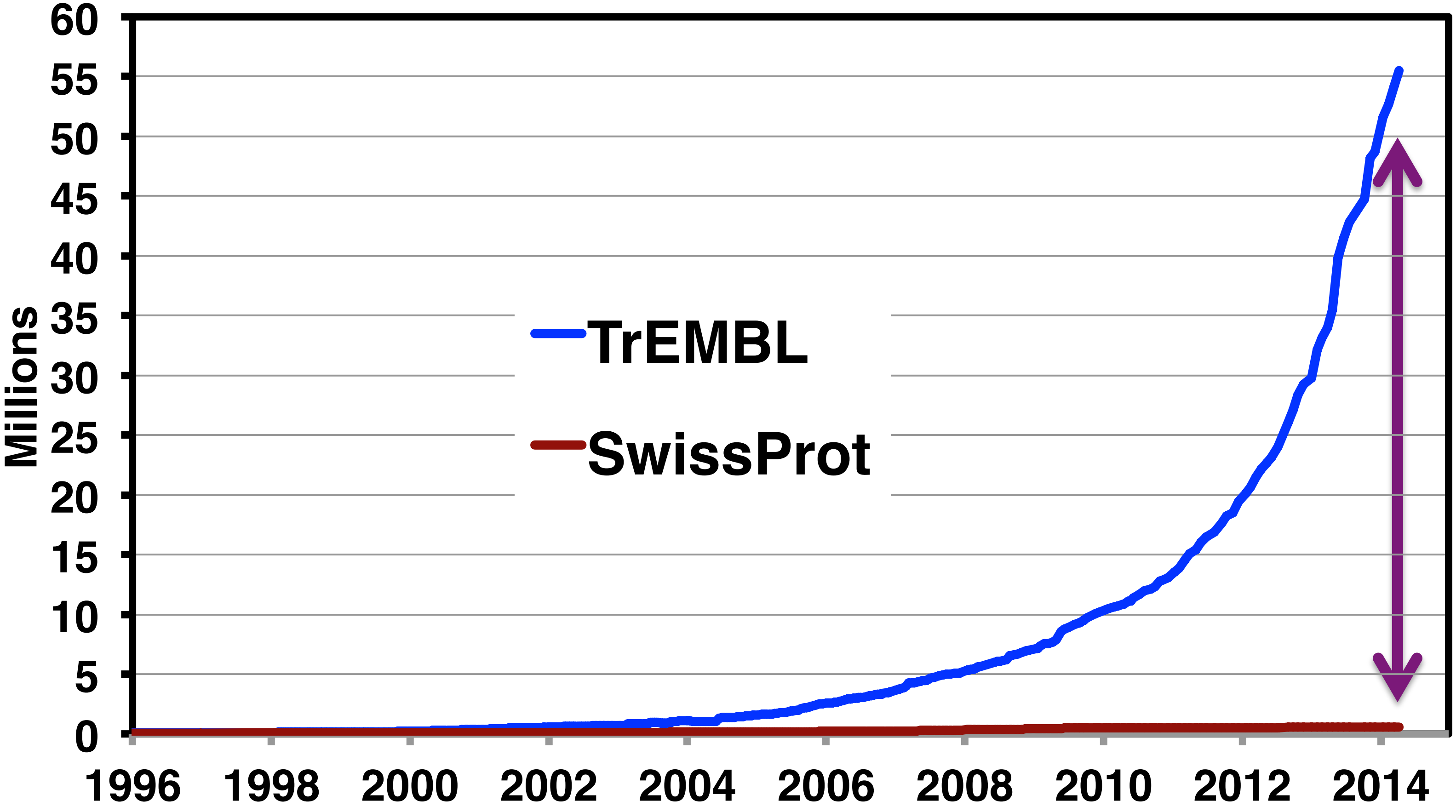
contains 544,996 sequence entries.



Perhaps 50% have unknown or uncertain functions

Release 2014_04 of 16-Apr-2014 of UniProtKB

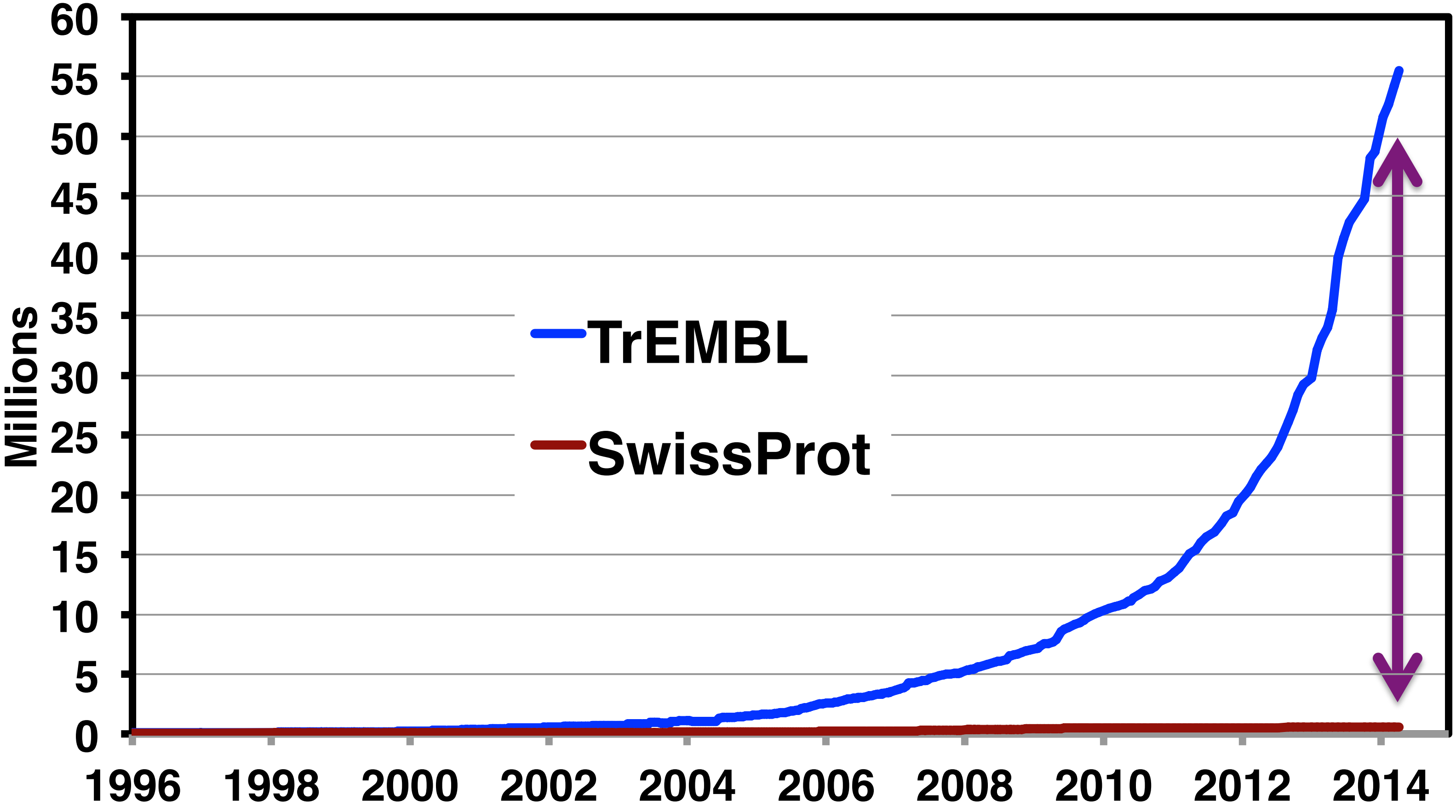
contains 55,503,547 sequence entries.



How do we solve this problem ?

Release 2014_04 of 16-Apr-2014 of UniProtKB

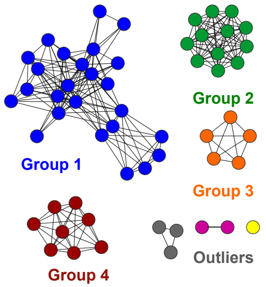
contains 55,503,547 sequence entries.



Requires high-throughput experimental and computational strategies

U54 GM093342: “Enzyme Function Initiative” (EFI)

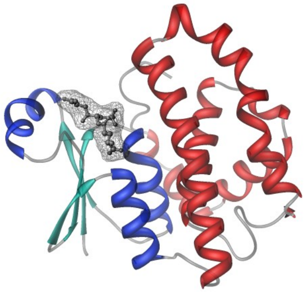
bioinformatics



sequence

»

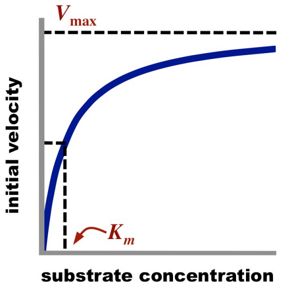
x-ray / computation



structure

»

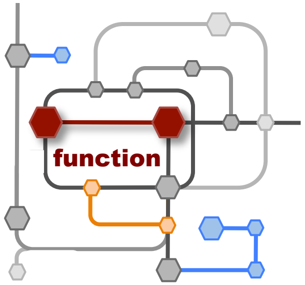
enzymology



reaction

»

biology



function

Albert Einstein
Steven Almo

Boston University
Karen Allen

Gladstone Institutes
Katherine Pollard

University of Illinois
John Gerlt
John Cronan
Jonathan Sweedler

UCSF
Matthew Jacobson
Andrej Sali
Brian Shoichet

University of New Mexico
Debra Dunaway-Mariano

Pennsylvania State
Squire Booker

University of Virginia
Wladek Minor

University of Utah
C. Dale Poulter

Deliverables, not Specific Aims

1. **Develop robust high-throughput sequence/structure-based tools and strategies** to discover *in vitro* activities and *in vivo* metabolic functions of unknown enzymes

2. **Disseminate tools to the community** for determining activities and functions of unknown enzymes

3. **Collaborate with the community** to implement the tools and strategies

4. **Correct annotations** in the protein databases

Deliverables, not Specific Aims

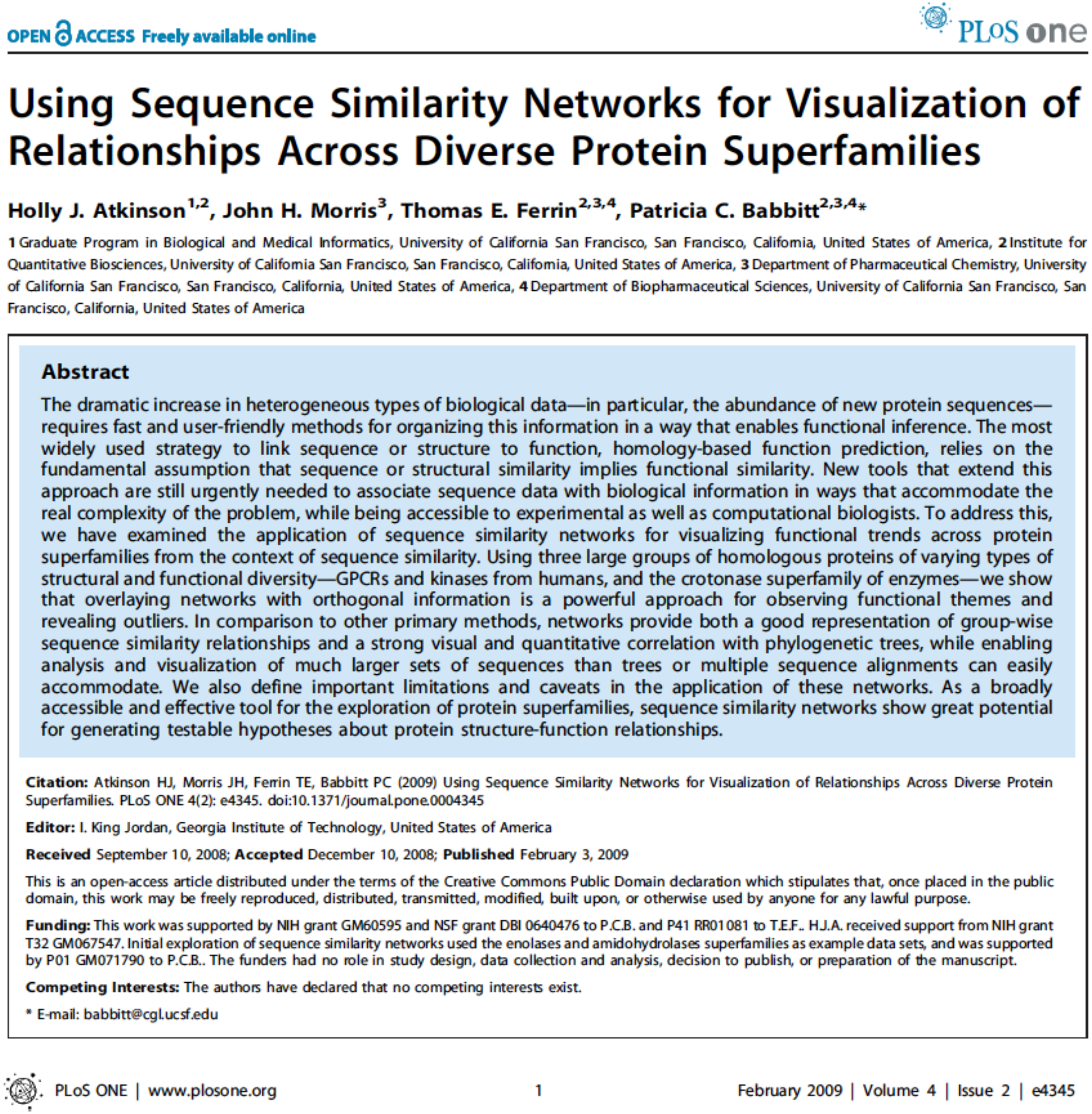
1. **Develop robust high-throughput sequence/structure-based tools and strategies** to discover *in vitro* activities and *in vivo* metabolic functions of unknown enzymes

2. **Disseminate tools to the community** for determining activities and functions of unknown enzymes

3. **Collaborate with the community** to implement the tools and strategies

4. **Correct annotations** in the protein databases

Sequence similarity networks: Atkinson et al.



OPEN ACCESS Freely available online

PLoS one

Using Sequence Similarity Networks for Visualization of Relationships Across Diverse Protein Superfamilies

Holly J. Atkinson[1,2], John H. Morris[3], Thomas E. Ferrin[2,3,4], Patricia C. Babbitt[2,3,4]*

1 Graduate Program in Biological and Medical Informatics, University of California San Francisco, San Francisco, California, United States of America, 2 Institute for Quantitative Biosciences, University of California San Francisco, San Francisco, California, United States of America, 3 Department of Pharmaceutical Chemistry, University of California San Francisco, San Francisco, California, United States of America, 4 Department of Biopharmaceutical Sciences, University of California San Francisco, San Francisco, California, United States of America

Abstract

The dramatic increase in heterogeneous types of biological data—in particular, the abundance of new protein sequences—requires fast and user-friendly methods for organizing this information in a way that enables functional inference. The most widely used strategy to link sequence or structure to function, homology-based function prediction, relies on the fundamental assumption that sequence or structural similarity implies functional similarity. New tools that extend this approach are still urgently needed to associate sequence data with biological information in ways that accommodate the real complexity of the problem, while being accessible to experimental as well as computational biologists. To address this, we have examined the application of sequence similarity networks for visualizing functional trends across protein superfamilies from the context of sequence similarity. Using three large groups of homologous proteins of varying types of structural and functional diversity—GPCRs and kinases from humans, and the crotonase superfamily of enzymes—we show that overlaying networks with orthogonal information is a powerful approach for observing functional themes and revealing outliers. In comparison to other primary methods, networks provide both a good representation of group-wise sequence similarity relationships and a strong visual and quantitative correlation with phylogenetic trees, while enabling analysis and visualization of much larger sets of sequences than trees or multiple sequence alignments can easily accommodate. We also define important limitations and caveats in the application of these networks. As a broadly accessible and effective tool for the exploration of protein superfamilies, sequence similarity networks show great potential for generating testable hypotheses about protein structure-function relationships.

Citation: Atkinson HJ, Morris JH, Ferrin TE, Babbitt PC (2009) Using Sequence Similarity Networks for Visualization of Relationships Across Diverse Protein Superfamilies. PLoS ONE 4(2): e4345. doi:10.1371/journal.pone.0004345

Editor: I. King Jordan, Georgia Institute of Technology, United States of America

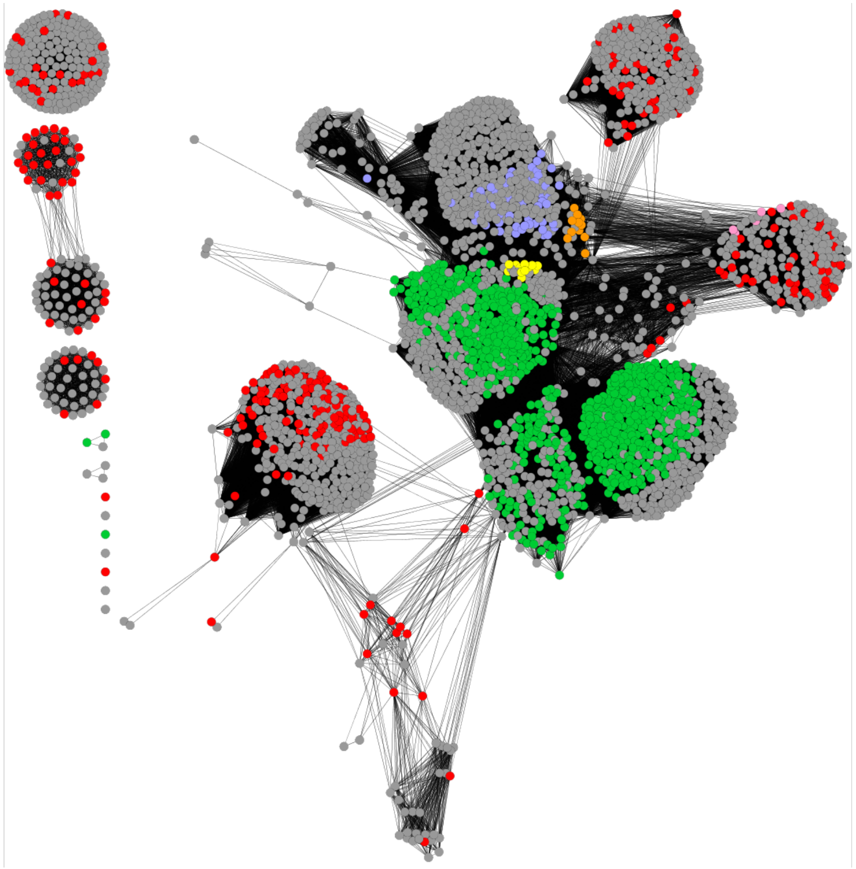
Received September 10, 2008; **Accepted** December 10, 2008; **Published** February 3, 2009

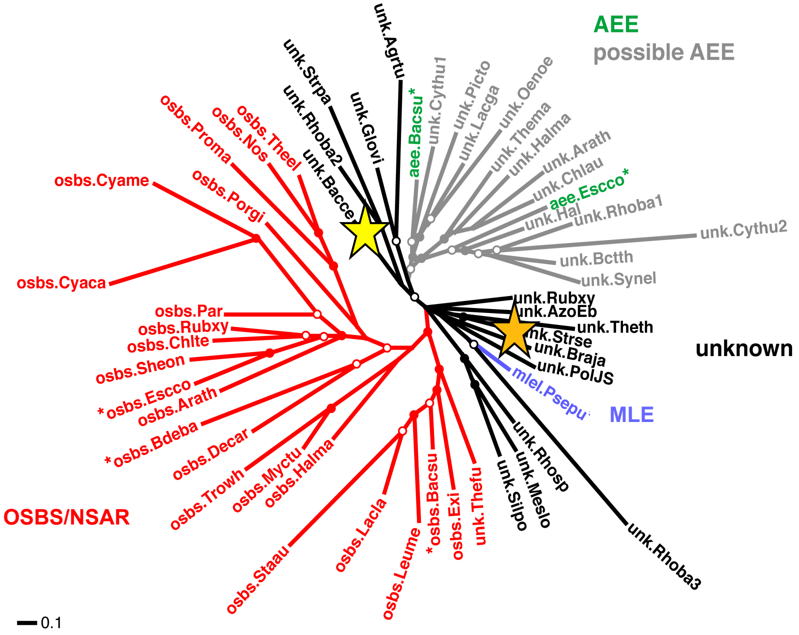
Funding: This work was supported by NIH grant GM60595 and NSF grant DBI 0640476 to P.C.B. and P41 RR01081 to T.E.F.. H.J.A. received support from NIH grant T32 GM067547. Initial exploration of sequence similarity networks used the enolases and amidohydrolases superfamilies as example data sets, and was supported by P01 GM071790 to P.C.B.. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: babbitt@cgl.ucsf.edu

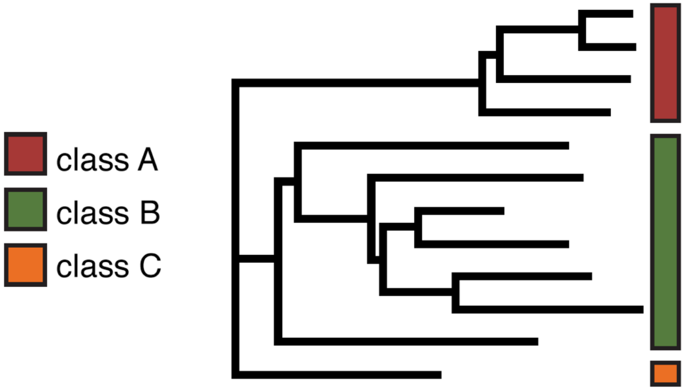
Sequence similarity networks (SSNs) vs dendrograms: enolase superfamily



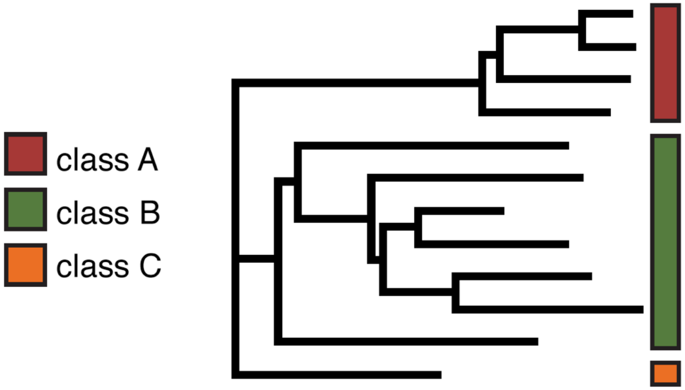


Families are easier to visualize in SSNs, so hypotheses are easier to formulate and explore

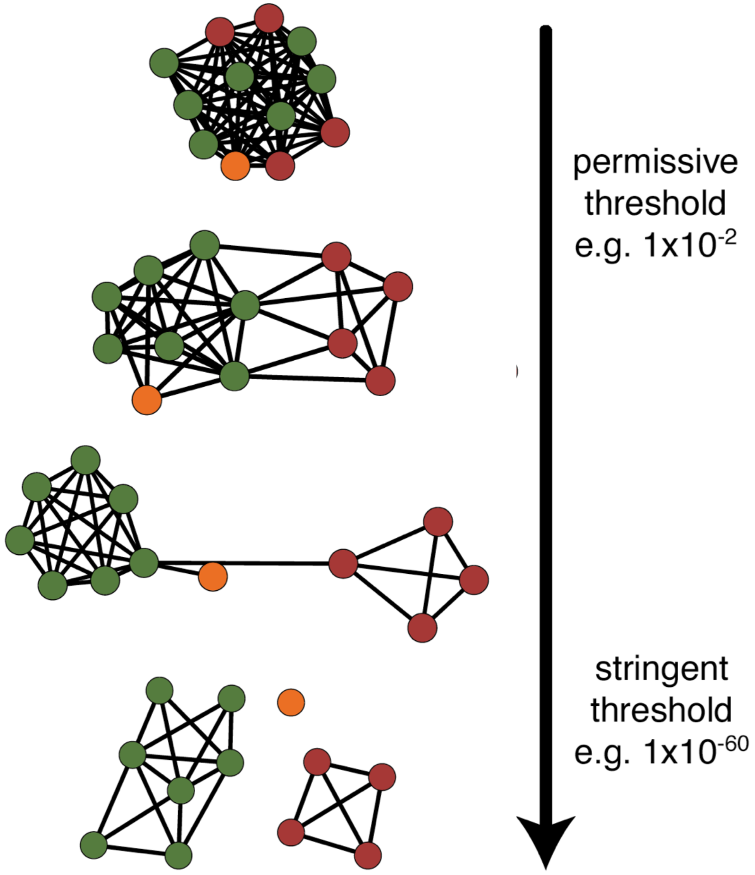
Dendrograms/trees for sequence relationships



Connectivity: multiple sequence alignments

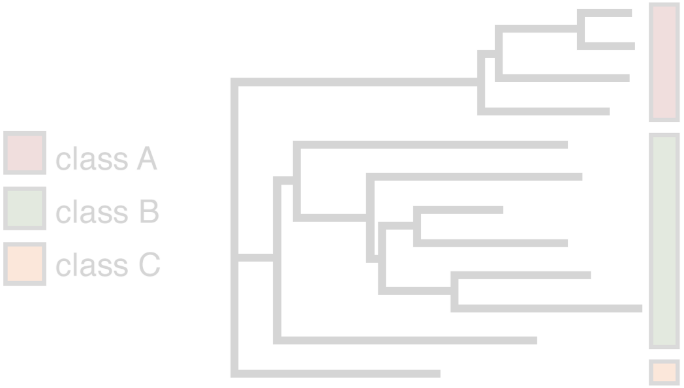


Sequence similarity networks



permissive threshold e.g. $1x10^{-2}$

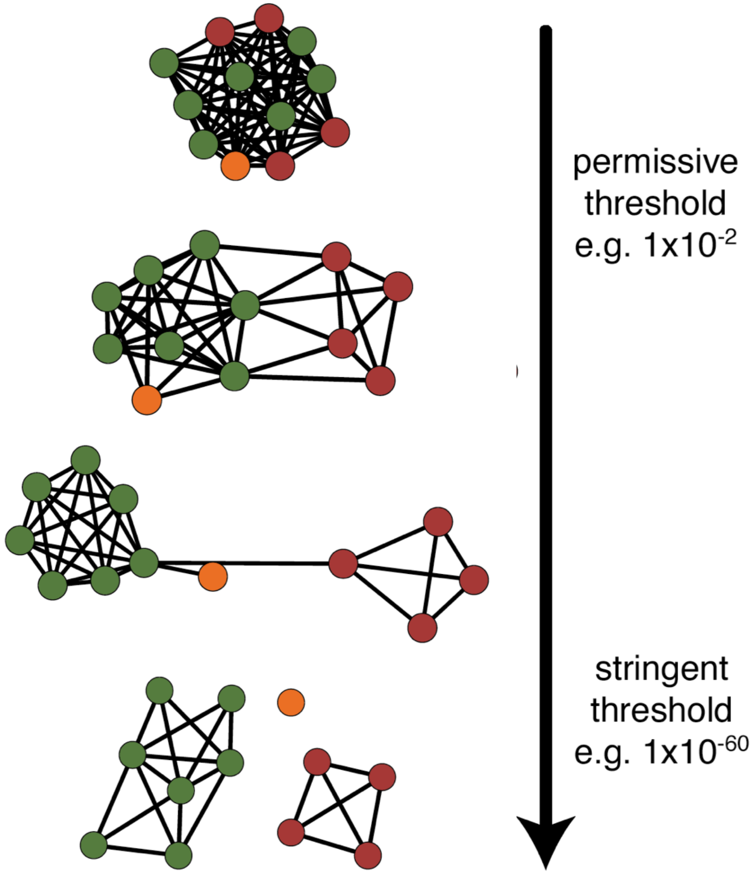
stringent threshold e.g. $1x10^{-60}$



node (circle) = sequence

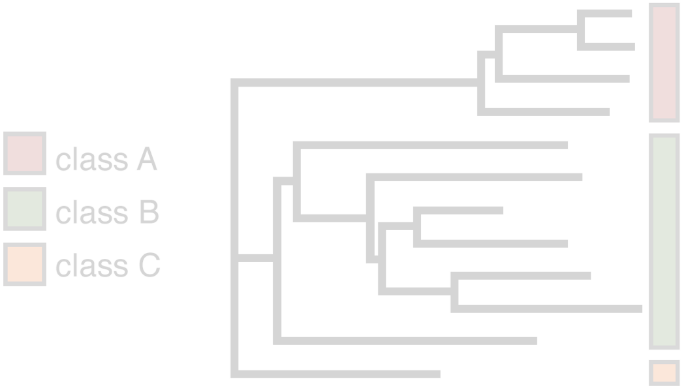
edge (line) = connection less than a user-defined score (e-value)

Connectivity: all-by-all BLASTP e-values



permissive threshold e.g. $1x10^{-2}$

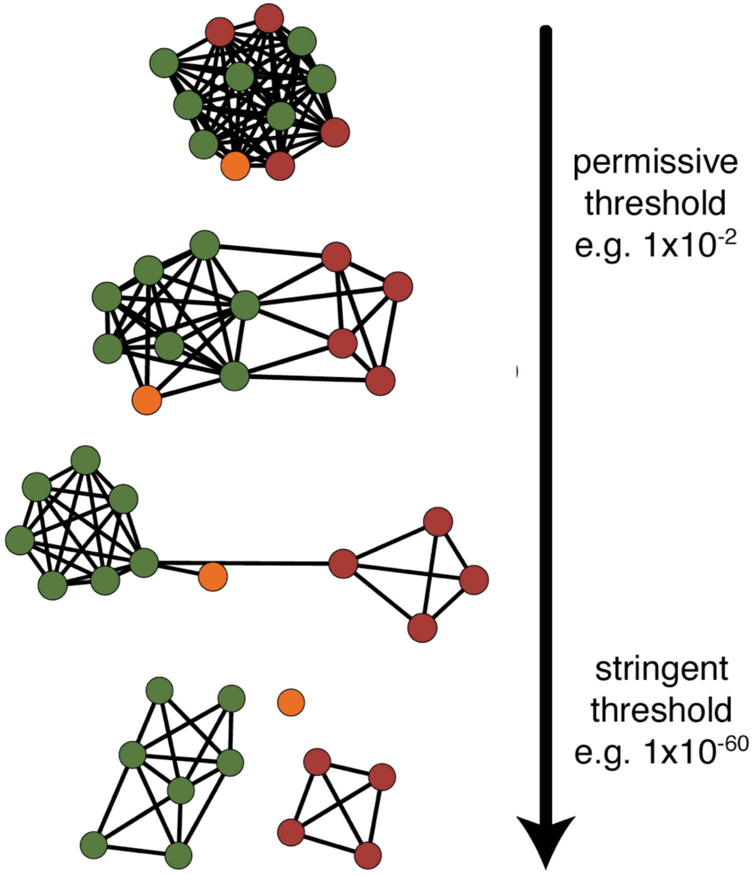
stringent threshold e.g. $1x10^{-60}$



node (circle) = sequence

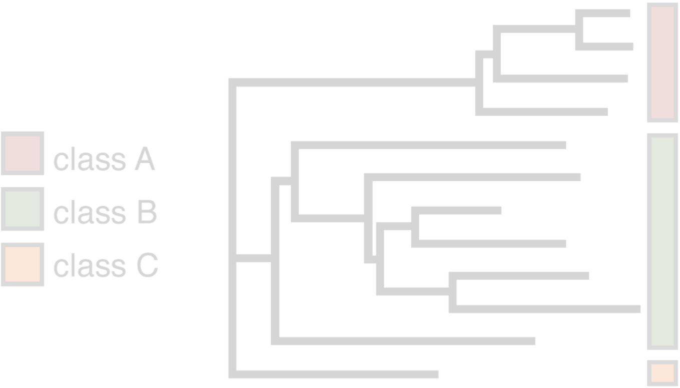
edge (line) = connection less than a user-defined score (e-value)

Faster to calculate than dendrograms



permissive threshold e.g. 1x10^{-2}

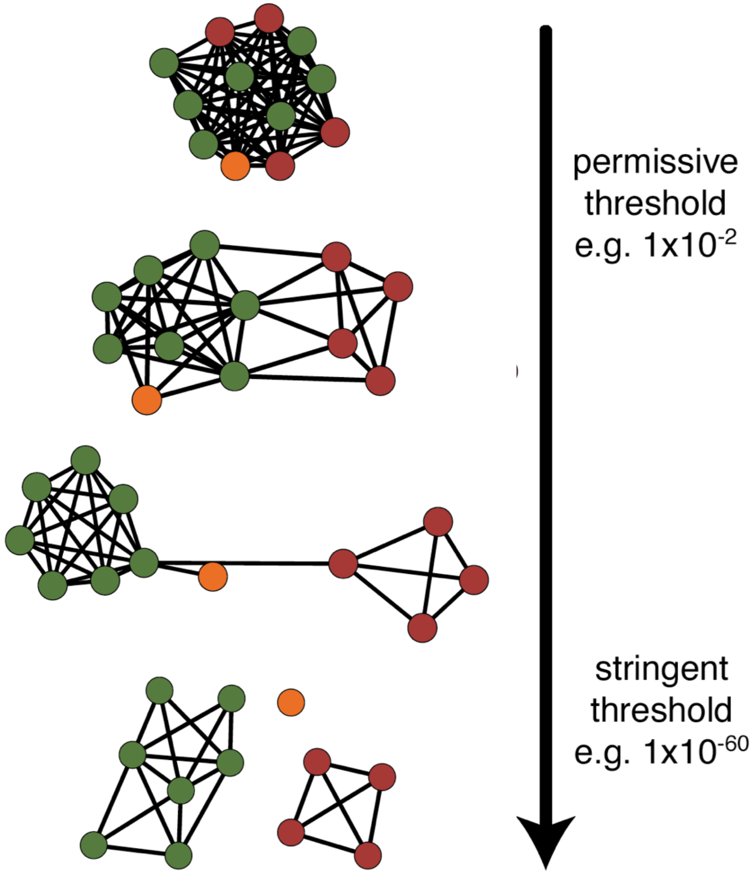
stringent threshold e.g. 1x10^{-60}



node (circle) = sequence

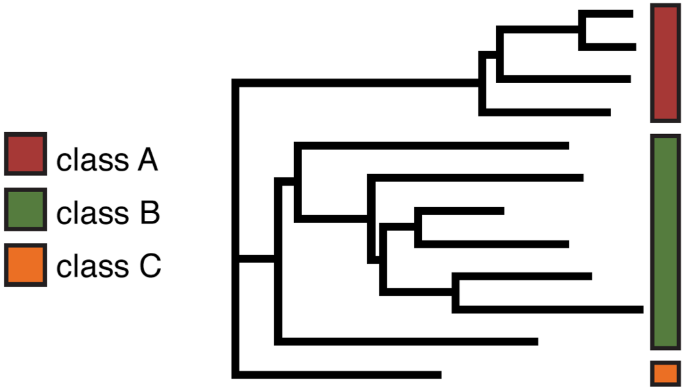
edge (line) = connection less than a user-defined score (e-value)

Qualitatively similar results



permissive threshold e.g. 1×10^{-2}

stringent threshold e.g. 1×10^{-60}

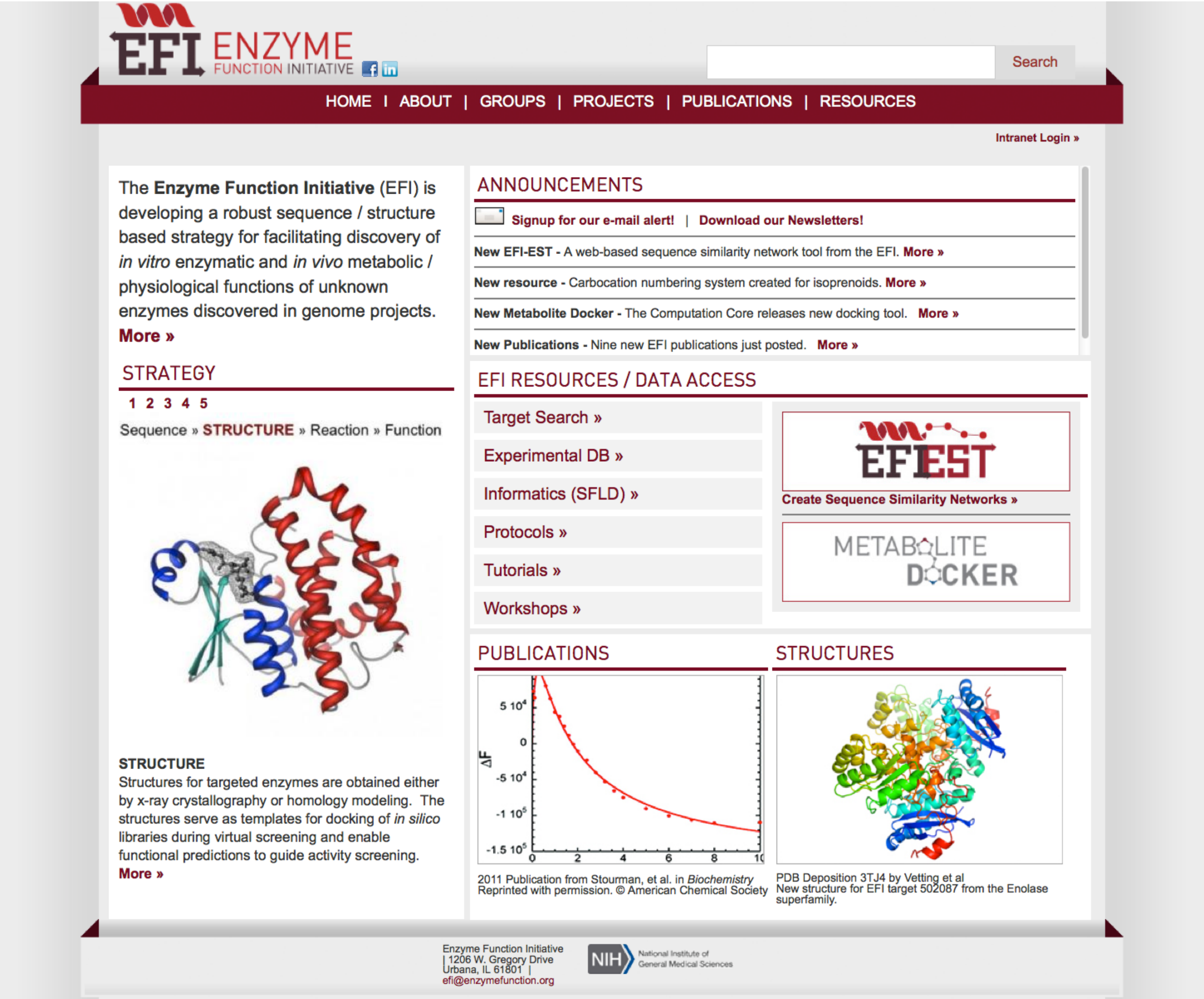


🟥 class A
🟩 class B
🟧 class C

node (circle) = sequence

edge (line) = connection less than a user-defined score (e-value)

EFI Home Page



EFI ENZYME FUNCTION INITIATIVE
Search
HOME | ABOUT | GROUPS | PROJECTS | PUBLICATIONS | RESOURCES
Intranet Login »

The **Enzyme Function Initiative** (EFI) is developing a robust sequence / structure based strategy for facilitating discovery of *in vitro* enzymatic and *in vivo* metabolic / physiological functions of unknown enzymes discovered in genome projects. **More »**

STRATEGY

1 2 3 4 5

Sequence » **STRUCTURE** » Reaction » Function

STRUCTURE
Structures for targeted enzymes are obtained either by x-ray crystallography or homology modeling. The structures serve as templates for docking of *in silico* libraries during virtual screening and enable functional predictions to guide activity screening.
More »

ANNOUNCEMENTS

Signup for our e-mail alert! | **Download our Newsletters!**

New EFI-EST - A web-based sequence similarity network tool from the EFI. **More »**

New resource - Carbocation numbering system created for isoprenoids. **More »**

New Metabolite Docker - The Computation Core releases new docking tool. **More »**

New Publications - Nine new EFI publications just posted. **More »**

EFI RESOURCES / DATA ACCESS

PUBLICATIONS

2011 Publication from Stourman, et al. in *Biochemistry*
Reprinted with permission. © American Chemical Society

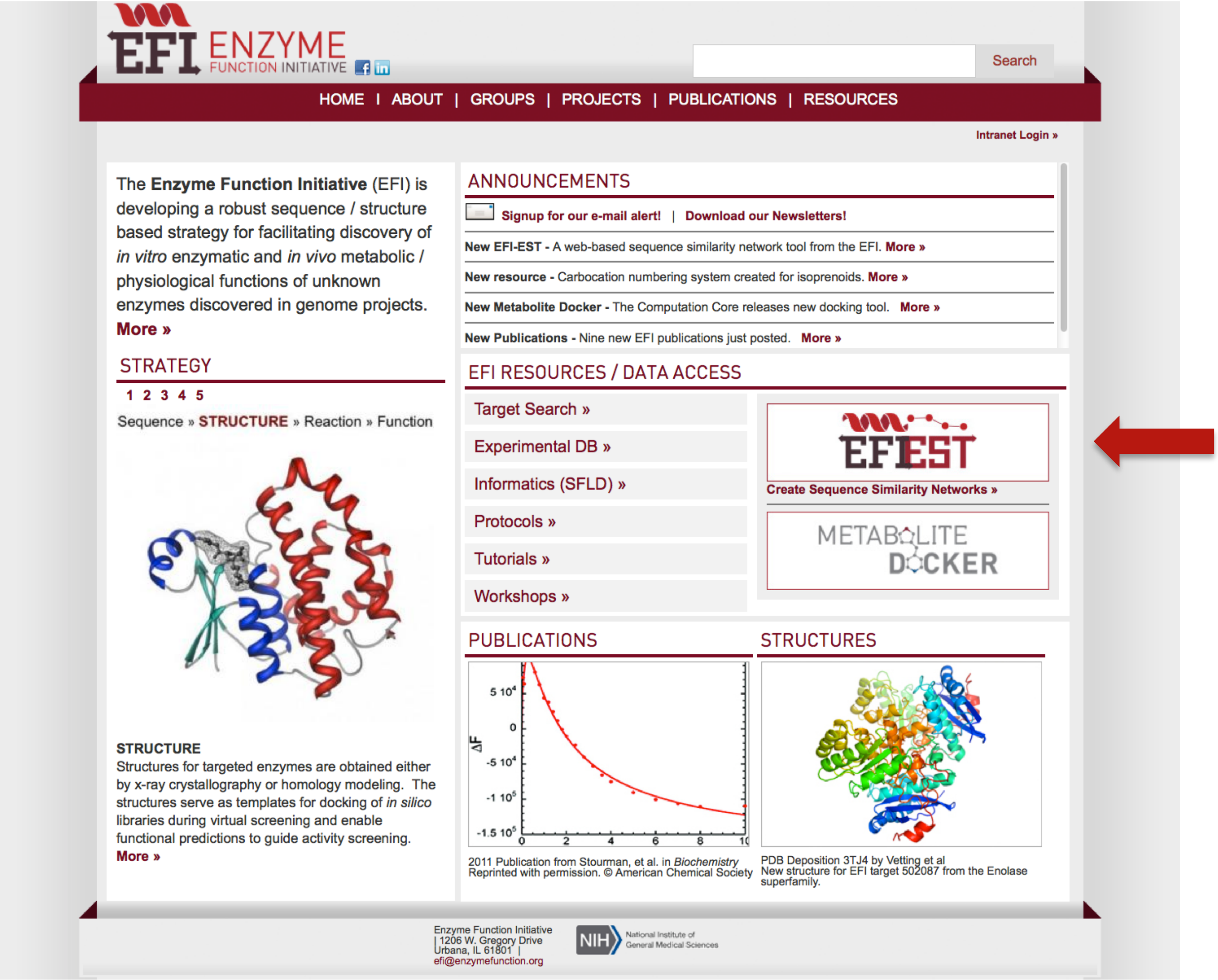
STRUCTURES

PDB Deposition 3TJ4 by Vetting et al
New structure for EFI target 502087 from the Enolase superfamily.

Enzyme Function Initiative
| 1206 W. Gregory Drive |
Urbana, IL 61801 |
efi@enzymefunction.org

http://enzymefunction.org/

EFI Home Page



EFI ENZYME FUNCTION INITIATIVE
Search
HOME | ABOUT | GROUPS | PROJECTS | PUBLICATIONS | RESOURCES
Intranet Login »

The **Enzyme Function Initiative** (EFI) is developing a robust sequence / structure based strategy for facilitating discovery of *in vitro* enzymatic and *in vivo* metabolic / physiological functions of unknown enzymes discovered in genome projects. **More »**

STRATEGY

1 2 3 4 5

Sequence » **STRUCTURE** » Reaction » Function

STRUCTURE
Structures for targeted enzymes are obtained either by x-ray crystallography or homology modeling. The structures serve as templates for docking of *in silico* libraries during virtual screening and enable functional predictions to guide activity screening.
More »

ANNOUNCEMENTS

Signup for our e-mail alert! | **Download our Newsletters!**

New EFI-EST - A web-based sequence similarity network tool from the EFI. **More »**

New resource - Carbocation numbering system created for isoprenoids. **More »**

New Metabolite Docker - The Computation Core releases new docking tool. **More »**

New Publications - Nine new EFI publications just posted. **More »**

EFI RESOURCES / DATA ACCESS

- Target Search »
- Experimental DB »
- Informatics (SFLD) »
- Protocols »
- Tutorials »
- Workshops »

Create Sequence Similarity Networks »

PUBLICATIONS

2011 Publication from Stourman, et al. in *Biochemistry*
Reprinted with permission. © American Chemical Society

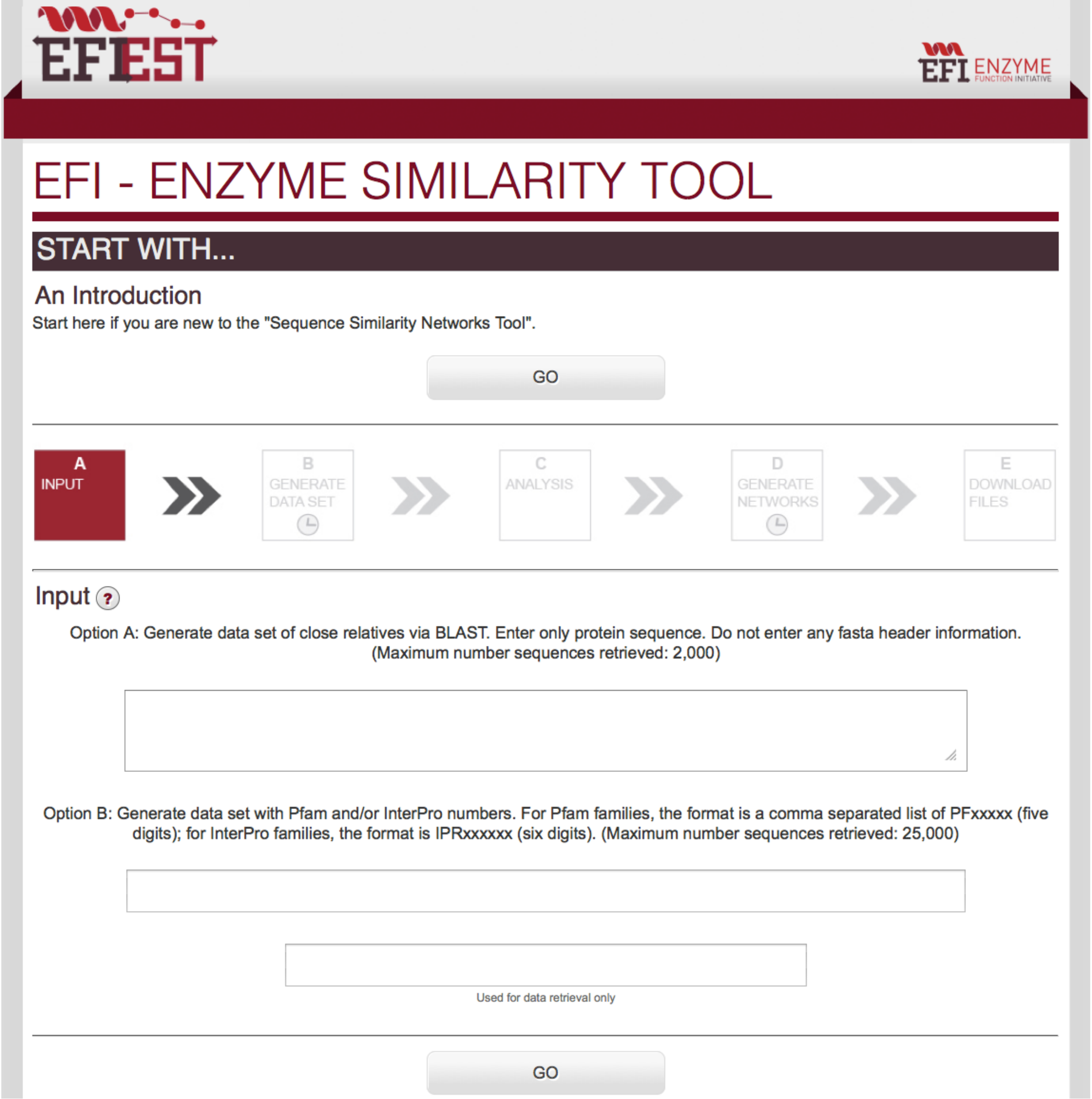
STRUCTURES

PDB Deposition 3TJ4 by Vetting et al
New structure for EFI target 502087 from the Enolase superfamily.

Enzyme Function Initiative
| 1206 W. Gregory Drive
Urbana, IL 61801 |
efi@enzymefunction.org

http://enzymefunction.org/

EFI-EST: user-initiated sequence similarity networks



EFI - ENZYME SIMILARITY TOOL

START WITH...

An Introduction

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

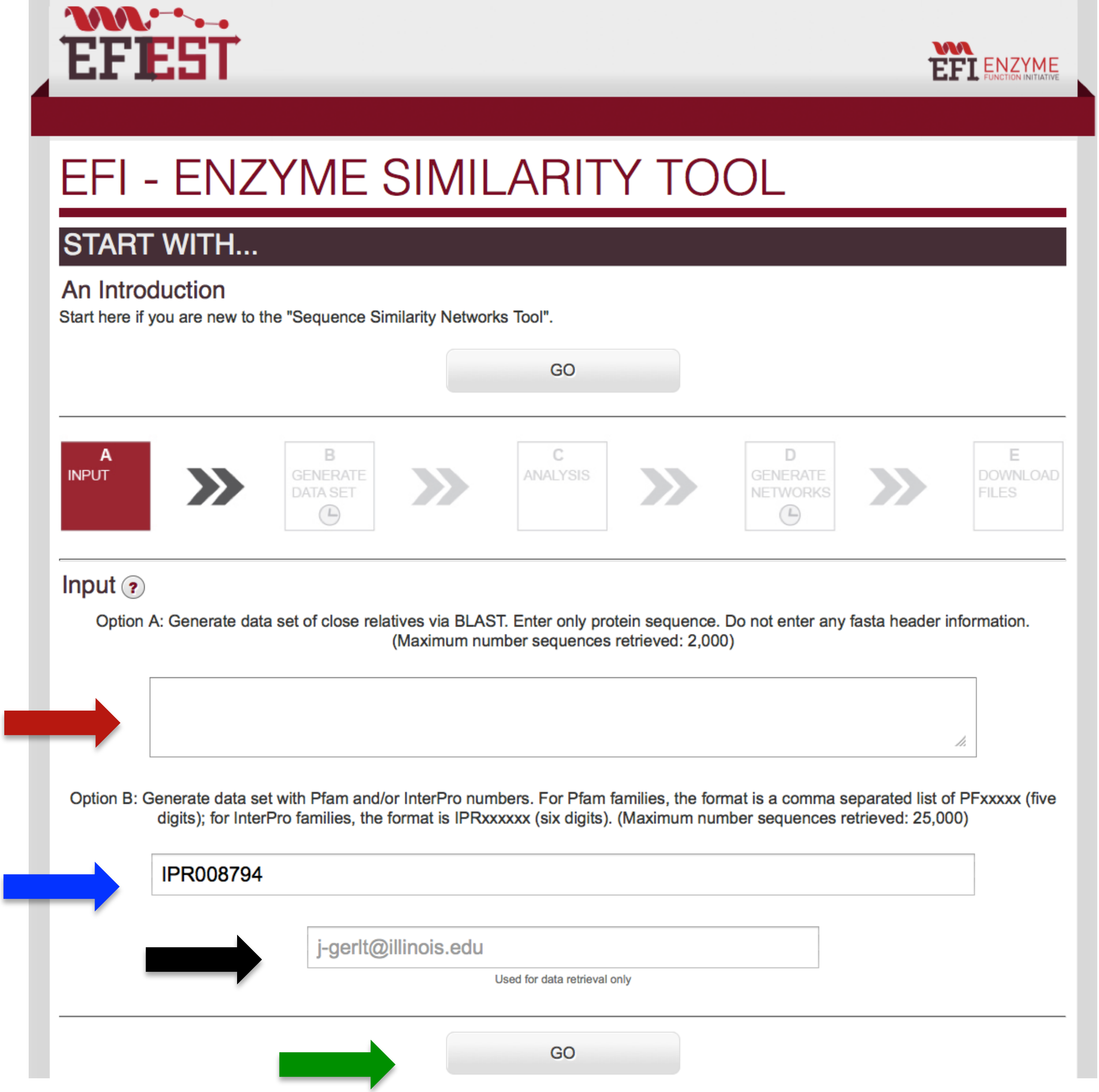
Input ?

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: 2,000)

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: 25,000)

Used for data retrieval only

Input sequence for BLAST or Pfam/InterPro family(ies)



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: 2,000)

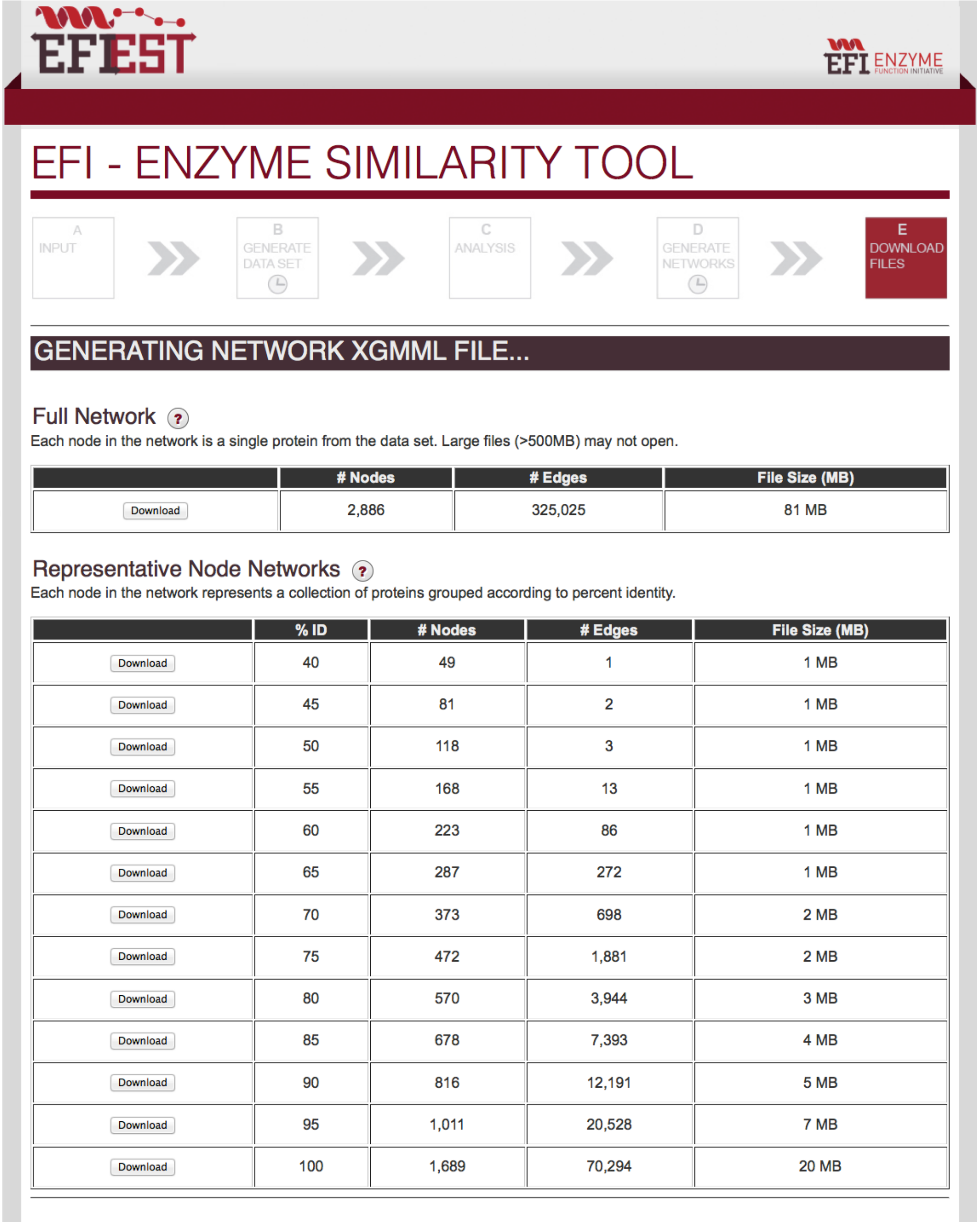
Option B: Generate data set with Pfam and/or InterPro numbers. For Pfam families, the format is a comma separated list of PFxxxxx (five digits); for InterPro families, the format is IPRxxxxxx (six digits). (Maximum number sequences retrieved: 25,000)

IPR008794

j-gerlt@illinois.edu

Used for data retrieval only

Output full and rep node networks (.xgmml files)



EFI - ENZYME SIMILARITY TOOL

GENERATING NETWORK XGMML FILE...

Full Network ?

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

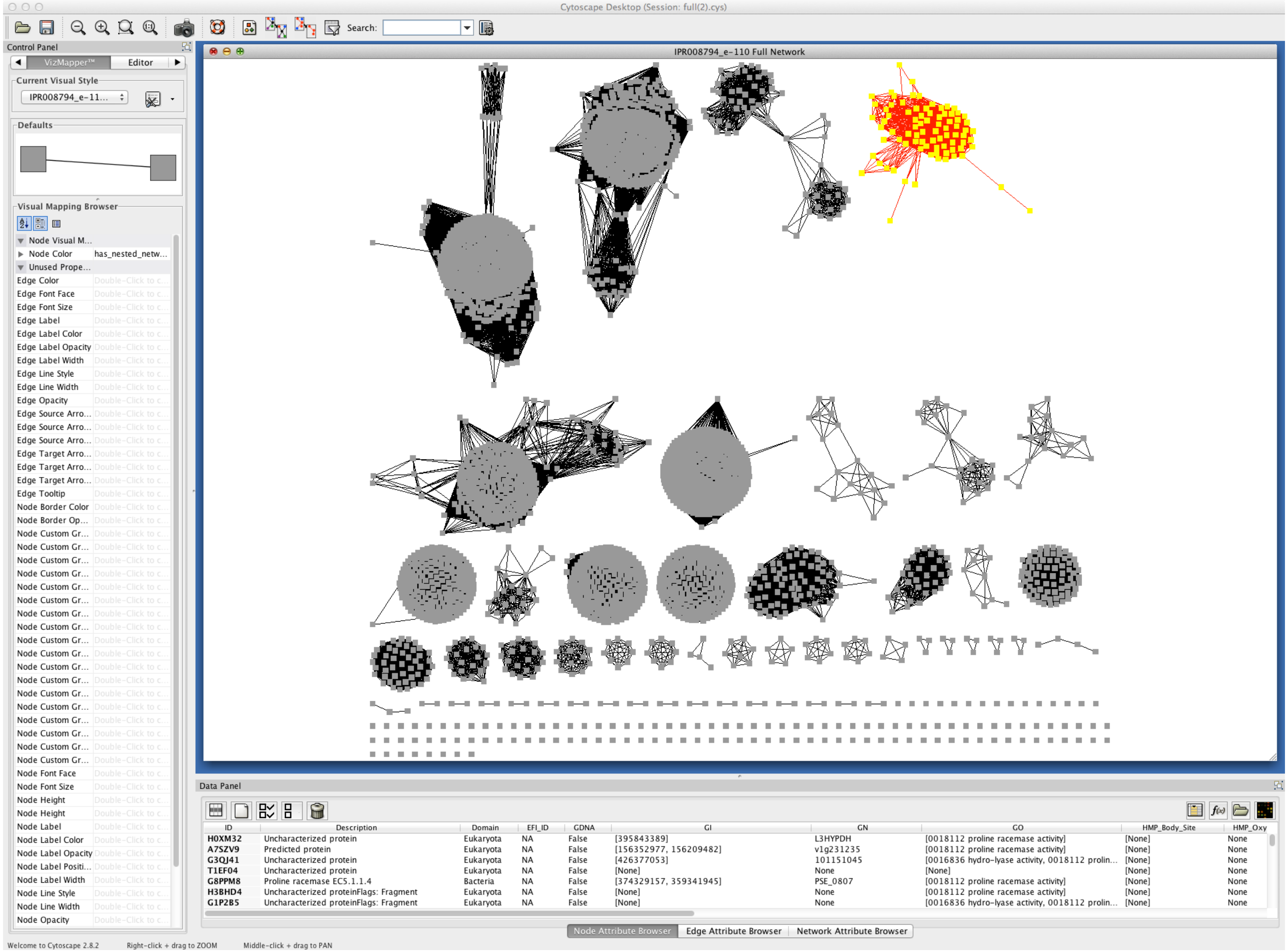
	# Nodes	# Edges	File Size (MB)
Download	2,886	325,025	81 MB

Representative Node Networks ?

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

	% ID	# Nodes	# Edges	File Size (MB)
Download	40	49	1	1 MB
Download	45	81	2	1 MB
Download	50	118	3	1 MB
Download	55	168	13	1 MB
Download	60	223	86	1 MB
Download	65	287	272	1 MB
Download	70	373	698	2 MB
Download	75	472	1,881	2 MB
Download	80	570	3,944	3 MB
Download	85	678	7,393	4 MB
Download	90	816	12,191	5 MB
Download	95	1,011	20,528	7 MB
Download	100	1,689	70,294	20 MB

Cytoscape 2.8.3: full network, e-110


Cytoscape Desktop (Session: full(2).cys)
Search:
Control Panel
VizMapper™
Editor
IPR008794_e-110 Full Network
Data Panel
Node Attribute Browser
Edge Attribute Browser
Network Attribute Browser
Welcome to Cytoscape 2.8.2
Right-click + drag to ZOOM
Middle-click + drag to PAN

Current Visual Style: IPR008794_e-11...

Defaults

Visual Mapping Browser

Node Visual M...	
Node Color	has_nested_netw...
Unused Prope...	
Edge Color	Double-Click to c...
Edge Font Face	Double-Click to c...
Edge Font Size	Double-Click to c...
Edge Label	Double-Click to c...
Edge Label Color	Double-Click to c...
Edge Label Opacity	Double-Click to c...
Edge Label Width	Double-Click to c...
Edge Line Style	Double-Click to c...
Edge Line Width	Double-Click to c...
Edge Opacity	Double-Click to c...
Edge Source Arro...	Double-Click to c...
Edge Source Arro...	Double-Click to c...
Edge Source Arro...	Double-Click to c...
Edge Target Arro...	Double-Click to c...
Edge Target Arro...	Double-Click to c...
Edge Target Arro...	Double-Click to c...
Edge Tooltip	Double-Click to c...
Node Border Color	Double-Click to c...
Node Border Op...	Double-Click to c...
Node Custom Gr...	Double-Click to c...
Node Custom Gr...	Double-Click to c...
Node Custom Gr...	Double-Click to c...
Node Custom Gr...	Double-Click to c...
Node Custom Gr...	Double-Click to c...
Node Custom Gr...	Double-Click to c...
Node Custom Gr...	Double-Click to c...
Node Custom Gr...	Double-Click to c...
Node Custom Gr...	Double-Click to c...
Node Custom Gr...	Double-Click to c...
Node Custom Gr...	Double-Click to c...
Node Custom Gr...	Double-Click to c...
Node Custom Gr...	Double-Click to c...
Node Custom Gr...	Double-Click to c...
Node Custom Gr...	Double-Click to c...
Node Custom Gr...	Double-Click to c...
Node Custom Gr...	Double-Click to c...
Node Custom Gr...	Double-Click to c...
Node Font Face	Double-Click to c...
Node Font Size	Double-Click to c...
Node Height	Double-Click to c...
Node Height	Double-Click to c...
Node Label	Double-Click to c...
Node Label Color	Double-Click to c...
Node Label Opacity	Double-Click to c...
Node Label Positi...	Double-Click to c...
Node Label Width	Double-Click to c...
Node Line Style	Double-Click to c...
Node Line Width	Double-Click to c...
Node Opacity	Double-Click to c...

ID	Description	Domain	EFI_ID	GDNA	GI	GN	GO	HMP_Body_Site	HMP_Oxy
H0XM32	Uncharacterized protein	Eukaryota	NA	False	[395843389]	L3HYPDH	[0018112 proline racemase activity]	[None]	None
A7SZV9	Predicted protein	Eukaryota	NA	False	[156352977, 156209482]	v1g231235	[0018112 proline racemase activity]	[None]	None
G3QJ41	Uncharacterized protein	Eukaryota	NA	False	[426377053]	101151045	[0016836 hydro-lyase activity, 0018112 prolin...	[None]	None
T1EF04	Uncharacterized protein	Eukaryota	NA	False	[None]	None	[None]	[None]	None
G8PPM8	Proline racemase EC5.1.1.4	Bacteria	NA	False	[374329157, 359341945]	PSE_0807	[0018112 proline racemase activity]	[None]	None
H3BHD4	Uncharacterized proteinFlags: Fragment	Eukaryota	NA	False	[None]	None	[0018112 proline racemase activity]	[None]	None
G1P2B5	Uncharacterized proteinFlags: Fragment	Eukaryota	NA	False	[None]	None	[0016836 hydro-lyase activity, 0018112 prolin...	[None]	None

Pre-computation of networks for Pfam families and clans

BLUE WATERS
SUSTAINED PETASCALE COMPUTING

User does not have to wait for BLASTs
Expedite hypotheses and experiments

EFI-EST uses protein families and clans from Pfam

Pfam

Families: conserved protein families based a seed alignment of representative sequences that is used to generate a profile hidden Markov model (HMM).

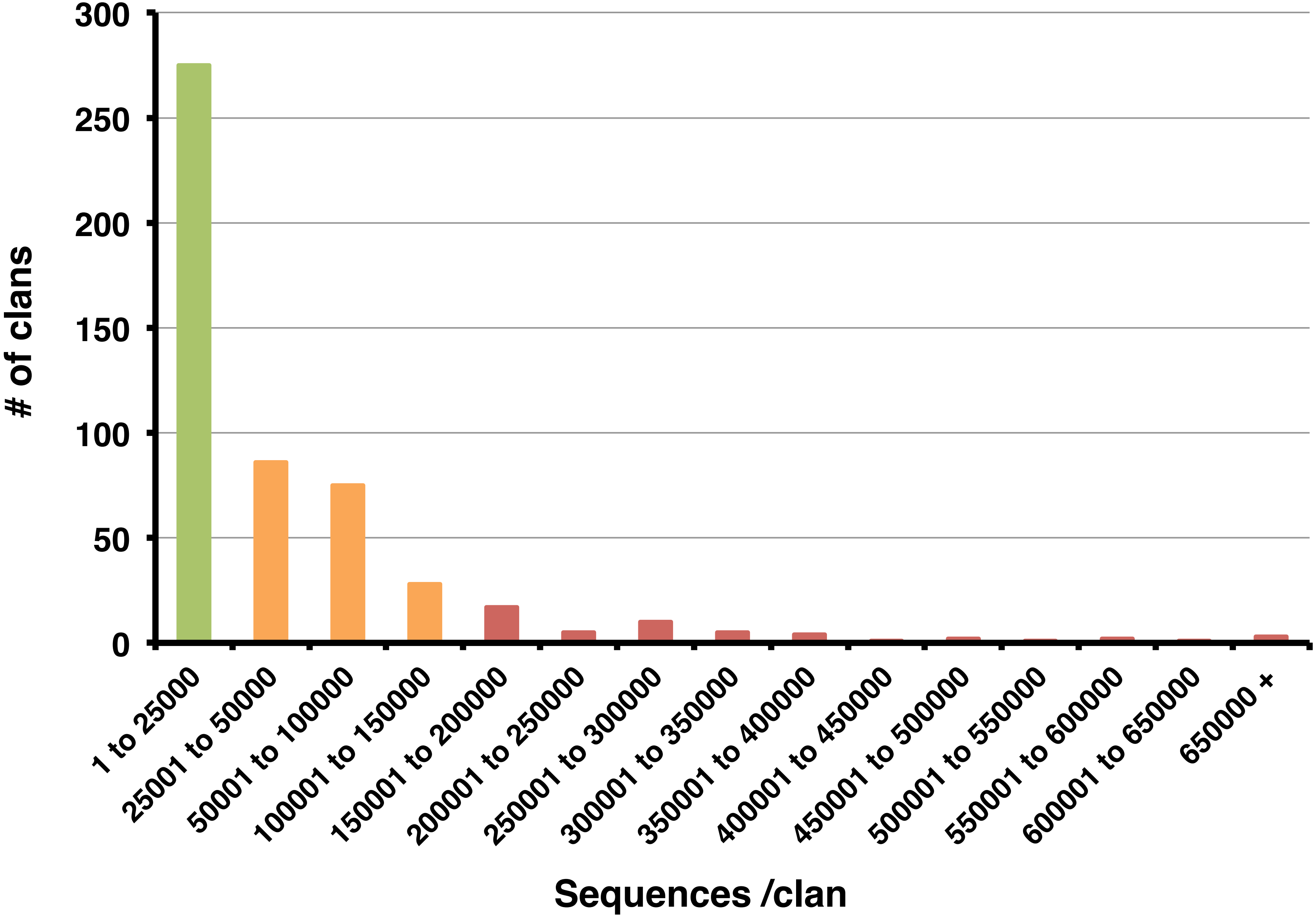
14,831 families in Pfam 27.0 (March 2013)

Clans: families (superfamilies) that have a common evolutionary ancestor based on structure and sequence.

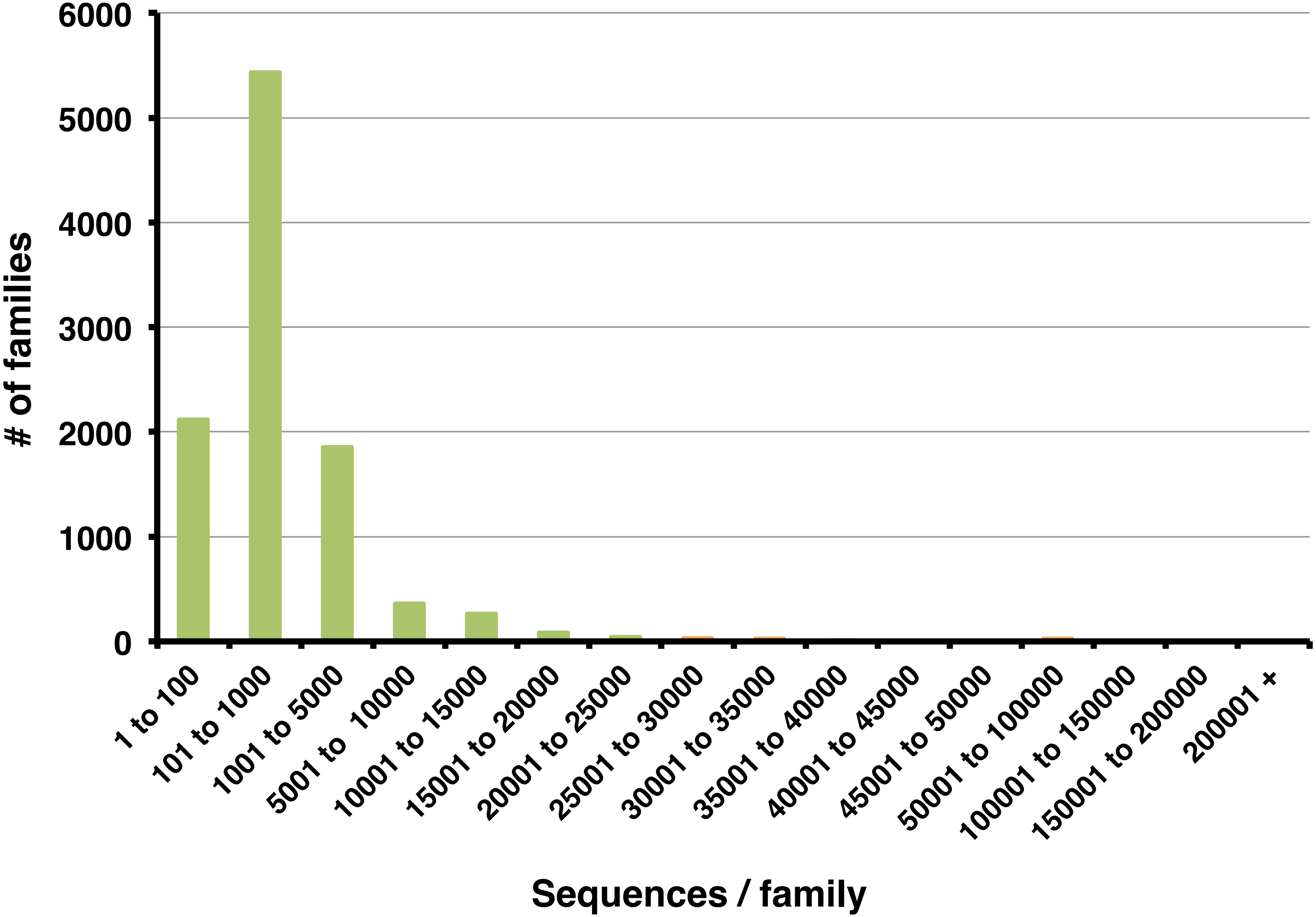
515 clans in Pfam 27.0 containing 4,563 Pfam families

http://pfam.sanger.ac.uk/

515 clans (4,563 families): 68,545 sequences/clan



10,268 “clanless”-families: 1,909 sequences/family



BLASTall on Blue Waters:
32 processors /node and 64 GB RAM/node

1. Partition each family into smaller sets of query sequences, e.g., 100 sequences for small families

2. Load the entire set of sequences into RAM

3. Use BLASTall to calculate e-values for the query sets against all sequences

4. Store BLAST results

5. Concatentate results

6. Filter results to remove redundancy (A-B vs B-A)

BLASTall on Blue Waters: 32 processors /node and 64 GB RAM/node

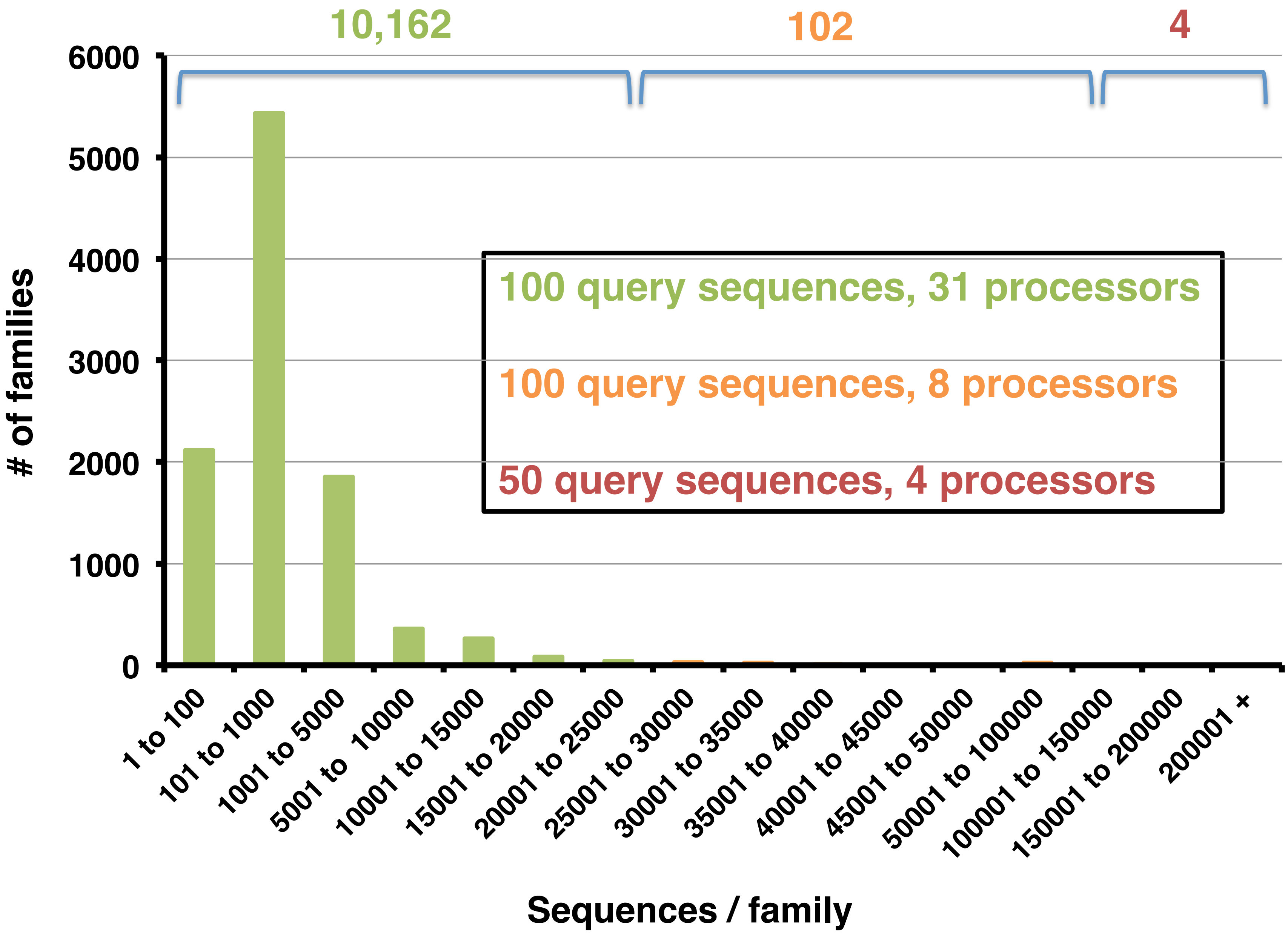
24 hr wall time limits the sizes of the query sets, depending on total number of sequences

Solution: decrease sequences into smaller query sets. Or, split family into smaller pieces.

64GB RAM limits the number of sequences that can be loaded and the number of results that can be stored

Solution: decrease number of processors to make more RAM available per processor but this increases the number of required nodes. Or, split family into smaller pieces. These are computationally equivalent.

10,268 families: average 1,909 sequences/family



100 query sequences, 31 processors

100 query sequences, 8 processors

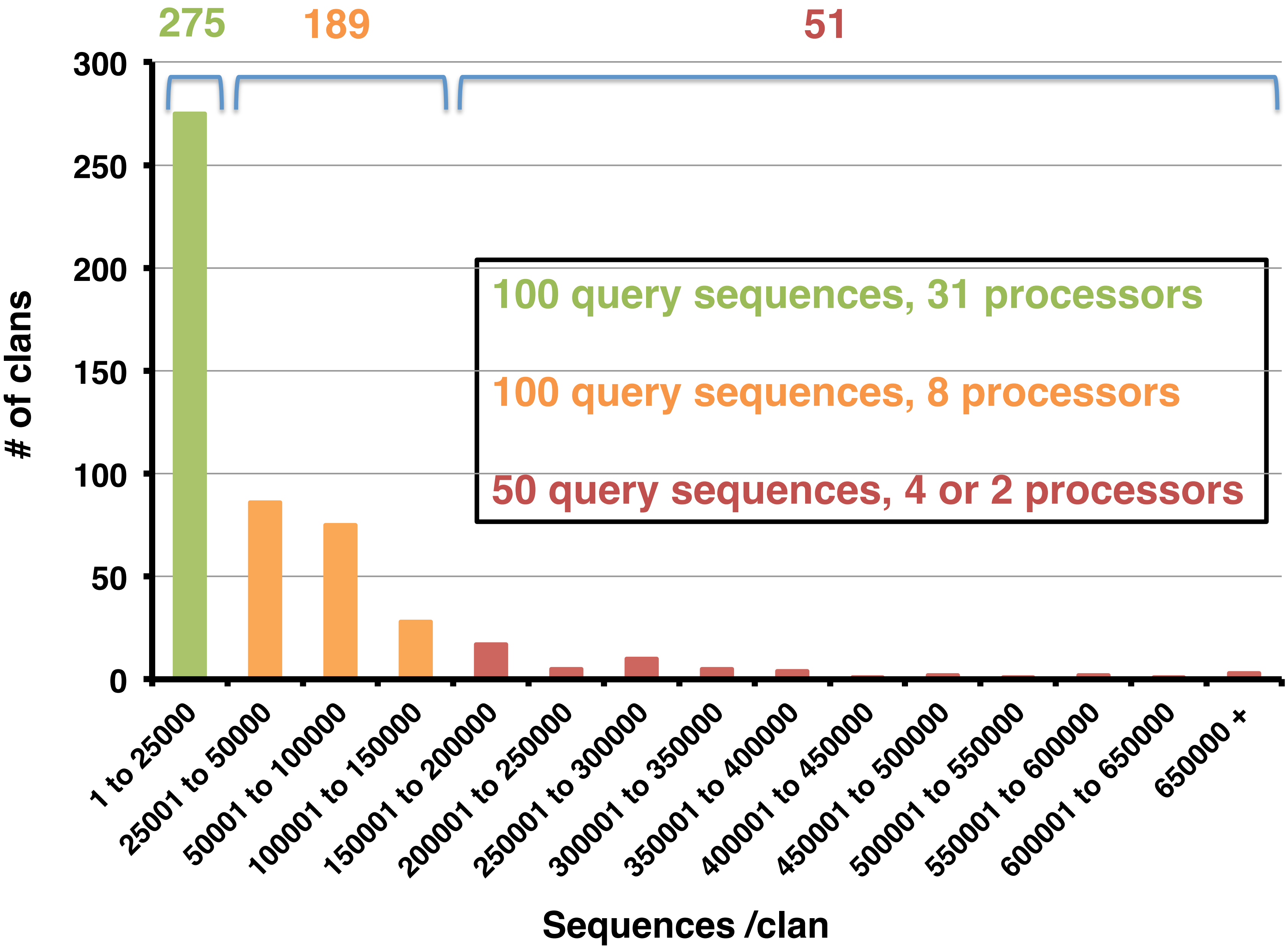
50 query sequences, 4 processors

10,268 “clanless”-families

To date, 10,264 BLASTs are complete!

These are being processed to yield .xgmml files for Cytoscape. Statistics are being calculated for choice of visualization thresholds.

515 clans: average 68,545 sequences/clan



100 query sequences, 31 processors

100 query sequences, 8 processors

50 query sequences, 4 or 2 processors

515 clans (and their 4,564 families)

514 BLASTs completed!

The successful BLASTs are being processed to yield .xgmml files for Cytoscape as well as statistics to choose visualization thresholds.

Families are being extracted from the 514 completed clans (3,049/4,563 to date); these are being processed to yield .xgmml files and statistics.

The largest clan (CL0023: 3,066,502 sequences) is too large for Blue Waters, using our current algorithms.

Future

The .xgmml files for 514 clans and all 14,831 families will be made available via a Web server.

Alternative BLAST approaches will be developed for CL0023.

Networks to be calculated on a quarterly refresh cycle to provide current networks to the biological community.

Personnel and Funding

University of Illinois, Urbana-Champaign

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

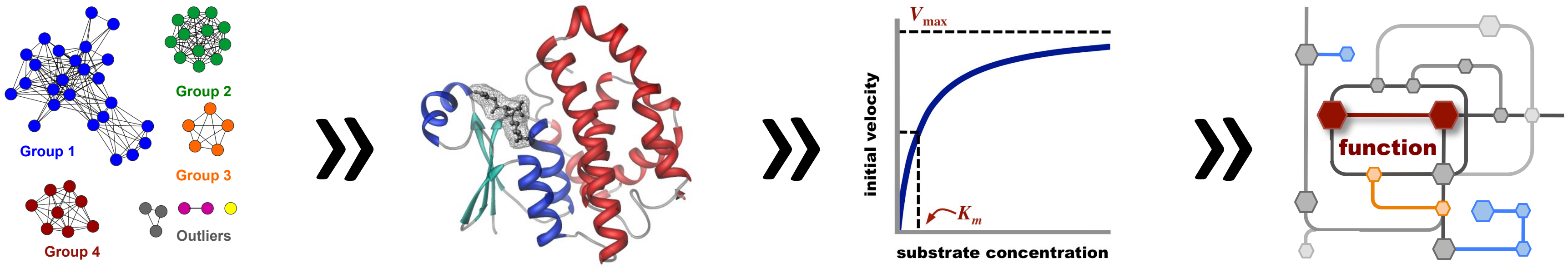
External Collaborators

Alex Bateman, EMBL-EBI
Matthew Jacobson, UCSF

NIH U54GM093342
UIUC Blue Waters Allocation

U54 GM093342: "Enzyme Function Initiative" (EFI)

bioinformatics *x-ray / computation* *enzymology* *biology*



sequence **structure** **reaction** **function**

[Albert Einstein](#)
Steven Almo

[Boston University](#)
Karen Allen

[Gladstone Institutes](#)
Katherine Pollard

[University of Illinois](#)
John Gerlt
John Cronan
Jonathan Sweedler

[UCSF](#)
Matthew Jacobson
Andrej Sali
Brian Shoichet

[University of New Mexico](#)
Debra Dunaway-Mariano

[Pennsylvania State](#)
Squire Booker

[University of Virginia](#)
Wladek Minor

[University of Utah](#)
C. Dale Poulter