Functional Annotation of the Protein Sequence Universe

Eugene Kolker

eugene.kolker@seattlechildrens.org
gnklkr@yahoo.com

Thanks to Jacek Becla and SLAC Team!
Outline

• Life sciences challenges and 4\textsuperscript{th} paradigm

• \textbf{Protein Functional Annotation:}
  – All vs. All BLAST: 10 mln vs. 10 mln proteins
  – Revitalizing COG db and clustering
  – Visualizing Protein Sequence Universe

• Ecosystem for Life sciences:
  – Data-Enabled Life Sciences Alliance (\textit{DELSA})
Past Grand Challenges & Solutions

• Green Revolution – Norman Borlaug
• Smallpox Eradication – William Foege
• Farmington Heart Study – Thomas Dawber

- Uber-goal, approach (data-driven, inter-disciplinary, technologies, collaborative), supportive ecosystem

Present Grand Challenges, Solutions and 4th paradigm of science – Jim Grey
Assigning functions to protein sequences

- Technologies produce huge data influx (e.g. mega-projects: 1000 genomes, EMP, i5K)
- 30% of proteins - unknown function
- Existing annotation databases overwhelmed
- Many no longer supported (e.g. COG, Systers, ClusTr)

- **Alignment (BLAST):** ~96% similar (identical)
  Jacek likes vanilla ice cream with chocolate on the top
  Jacek loves vanilla ice cream with chocolate on the top

- **Function Assignment**
  If the sequences are similar, their functions are similar too
COG- Clusters of Orthologous Groups (of proteins)

- Developed by Eugene Koonin, David Lipman and NCBI Team – over 4500 citations
- Classification of proteins into groups with common function
- Prokaryotes (COG): 66 genomes, 200K proteins, 5K clusters
- Eukaryotes (KOG): 7 genomes, 113K proteins, 5K clusters
Our approach

• Protein sequence databases grew from ~1 mln sequences to ~10 mln sequences (2010) and now ~14 mln sequences

• Thanks to Roger Barga and Microsoft Research Team: BLAST aligned All vs All (10 mln vs 10 mln) sequences and generated 3 filtered billion records

• Our Team (Roger Higdon): developed a robust alignment model based on length normalized bit score (LNBS)
COGs and Protein functional groups (PFGs)

UniRef100
9.9E6

- COG DB
  +3.5E6
- Bacteria & Archea
  5.3E6
- Rest
  1.9E6
- Eukaryotes
  2.8E6
- KOG DB
  +1.4E6

- left2cluster
  1.8E6
  SLC
  509,382 PFGs

- left2cluster
  1.5E6
  SLC
  105,671 PFGs

LNBS>4.3
LNBS>5.2
Hadoop and Clustering

• Clustering
  – Our Team: Natali Kolker and Bill Broomall
  – Implemented Hadoop/Hive to handle the data
  – Thanks to the Hadoop community
  – Use COG db as starting point
    LNBS >4.3: SENS & SPEC ~98%
  – COGs expended ~30-fold
  – Cluster remaining 1.8M proteins into PFGs by single linkage clustering
- Tested results by Amazon Elastic MapReduce
Single linkage clustering on Hadoop

COG

KOG
Further COG development

- Working with NCBI
- Using psiBLAST instead of BLAST
- Evaluating the modified approach
  - Bacterial genomes
  - COG test/train data
  - Compare to all vs. all BLAST approach
- Assign new proteins to COG clusters
- Cluster unassigned proteins
Protein Sequence Universe

• Visualization and analysis framework
  – Explore the relationship between proteins
• Multidimensional Scaling
  – Project sequence data into 3D
• Analyze and visualize protein data
• Benefit: Expand universe without All vs All
  - Interpolation
• Thanks to Geoffrey Fox and Indiana University Team
COG Protein Universe
Close-up of related COGs
Why do YOU care?

• This protein universe is inside and outside you!

• President/Office of Science, Technology and Policy’s Request on 21st Century BIOECONOMY - to meet grand challenges in health, food, energy, and environment in lean budget times.
Why DELSA Now? What is DELSA?

**Data-Enabled Life Sciences Alliance**

- The research necessity of the LS community to work across diverse domains and with computer, data and cyberinfrastructure experts to leverage LS opportunities
- Scientific progress and accelerated rate of LS result in a pressing need for reproducibility
- A perceived gap between the needs of data-enabled LS and current funding initiatives
- An urgent need to create supporting ecosystem of industry, academia, federal agencies, and foundations

**DELSA includes:**

Bio-sciences, (global) health, ecology, environment, (gen)omics, evolution, computer sciences, policies, cyberinfrastructure, clinical research, management.
DELSA: Vision and Mission

- Proposed in May 2011 at the NSF’s DISW2 workshop
- Next meeting: Nov. 13, 2011, SuperComputing SC11, Seattle

**DELSA Vision:**

Sustainable and shared access to data, knowledge, tools, and services to find solutions for the pressing needs of our global society. DELSA will advance data-enabled life sciences by moving from “one scientist-one project” to “collective innovation”.

**DELSA Mission:**

The mission of DELSA is to become the leading voice and coordinating framework to accelerate data-enabled research in the life sciences community.
Team: Beth Stewart  
Bill Broomall  
Carey Sheu  
Courtney MacNealy-Koch  
Larissa Stanberry  

Collaborators: Ken Zaret  
Gerald van Belle  
Gordon Cohen  
Michael Portman  
Roger Barga  

Natali Kolker  
Olga Castaner  
Roger Higdon  
Dean Welch  
Winn Haynes  

Support: NSF, NIH, McMillen Foundation  
Jim Hendricks  
Tom Hansen  

Thanks!
Summary

• Life sciences challenges and 4\textsuperscript{th} paradigm

• Protein Functional Annotation:
  – All vs. All BLAST: 10 mln vs. 10 mln proteins
  – Revitalizing COG db and clustering
  – Visualizing Protein Sequence Universe

• Ecosystem for Life sciences:
  – Data-Enabled Life Sciences Alliance (DELSA)
Get Involved

Life scientists, engineers, computer scientists, cyberinfrastructure, data and analysis experts, companies & institutions:

• Get involved in **DELSA**
• Participate in building supporting ecosystem
• Attend our workshop @SC11, Nov. 13: [www.delsall.org](http://www.delsall.org)
• Contact: eugene.kolker@seattlechildrens.org
gnklkr@yahoo.com

**Chief Data Officer** at Seattle Children’s: predictive analytics