VISUALIZING THE PROTEIN SEQUENCE UNIVERSE

L.STANBERRY¹, R.HIGDON¹, W.HAYNES¹, N.KOLKER¹, W.BROOMALL¹, S.EKANAYAKE², A.HUGHES², Y.RUAN², J.QIU², E.KOLKER¹, G.FOX²

¹SEATTLE CHILDREN’S, ²INDIANA UNIVERSITY

ECMLS 2012, 18 June 2012
A 4th paradigm problem in biology
- Assigning functions (annotating) proteins
- Challenge
- Our goal
- PSU: methods & initial results
- Conclusions
Grand Challenge of Functional Genomics

- New technologies produce peta- and exabytes of data
- Protein Sequence Universe (PSU), the protein sequence space, expand exponentially
  - EMP, i5K, iPlant, NEON
  - 30% of existing sequenced proteins unannotated
- Existing resources overwhelmed, many unsupported: COG, Systers, ClusTr, eggNOG.
Ultimate Goal: Annotate All Proteins

Our approach:

- Revitalize, expand & enhance protein annotation resources.
- Develop sustainable software framework.
- Use HPC and most powerful CI – grids & clouds.
- Provide rigorous and reliable tools to annotate protein sequences.
COG: Clusters of Orthologous Groups

- COG database was developed by NCBI.
- Proteins classified into groups with common function encoded in complete genomes.
- Prokaryotes (COG): 66 genomes, 200K proteins, 5K clusters.
- Eukaryotes (KOG): 7 genomes, 113K proteins, 5K clusters.
- Valuable scientific resource: 5K citations.
Clustering 10 million UniRef100

- UniRef100: 10M proteins including 5.3M bacterial & archaeal
- BLAST - common sequence alignment approach
- All vs. All alignment on Azure
  - 475 eight-core virtual machines produced 3+ billion filtered records in 6 days
Clustering 10 million UniRef100

- Use prokaryotic COG as a starting point.
- Expand COGs ~20 fold (3.5 million proteins).
- Cluster 2M proteins into 500K functional groups
  - Single linkage clustering with MapReduce framework on Hadoop
Promise and Challenge of Annotation

- Clustering facilitates mass annotation
  BUT
- Takes considerable efforts and expertise
- Multiple cloud systems and compute solutions
Public Resources

- Struggle to cope with the influx of data
- Provide limited interactive and analytic capabilities
- Many no longer supported (SYSTERS, CluSTr, COG)

- Biological community needs scalable, sustainable and efficient approach to visualize, explore and annotate new data.
Protein Sequence Universe

- PSU Goal: Enhance annotation resources with analytic and visualization (browser) tools.
- Project sequence data into 3D using multidimensional scaling (MDS).
  - MDS interpolation allows expanding the universe without time consuming all vs all $O(N^2)$
    - 3D map allows much faster interpolation
Multi-Dimensional Scaling (MDS)

- Sammon’s objective function

\[ H = \sum_{i<j}^{n} \frac{(f(\delta_{ij}) - d(x_i, x_j))^2}{f(\delta_{ij})} \]

- \( \delta_{ij} \) is dissimilarity measure between sequences \( i \) and \( j \)
- \( d \) is Euclidean distance between projections \( x_i \) and \( x_j \)
- Denominator: larger contribution from smaller dissimilarities
- \( f \) is monotone transformation of dissimilarity measure chosen “artistically”
Typical Metagenomic MDS

16S rRNA Random Sample of 100K Sequences Colored by Megaregion
MDS Details

- \( f \) chosen heuristically to increase the ratio of standard deviation to mean for \( f(\delta_{ij}) \) and to increase the range of dissimilarity measures.

- \( O(n^2) \) complexity to map \( n \) sequences into 3D.

- MDS can be solved using EM (SMACOF – fastest but limited) or directly by Newton's method (it’s just \( \chi^2 \)).

- Used robust implementation of nonlinear \( \chi^2 \) minimization with Levenberg-Marquardt.

- 3D projections visualized in PlotViz.
MDS Details

- Input Data: 100K sequences from well-characterized prokaryotic COGs.
- Proximity measure: sequence alignment scores
- Scores calculated using Needleman-Wunsch
- Scores “sqrt 4D” transformed and fed into MDS
  - Analytic form for transformation to 4D
  - $\delta_{ij}^n$ decreases dimension $n > 1$; increases $n < 1$
  - “sqrt 4D” reduced dimension of distance data from 244 for $\delta_{ij}$ to 14 for $f(\delta_{ij})$
  - Hence more uniform coverage of Euclidean space
3D View of 100K COG Sequences
Implementation

- NW computed in parallel on 100 node 8-core system.
- Used Twister (IU) in the Reduce phase of MapReduce
- MDS Calculations performed on 768 core MS HPC cluster (32 nodes)
- Scaling, parallel MPI with threading intranode
- Parallel efficiency of the code approximately 70%
- Lost efficiency due memory bandwidth saturation
- NW required 1 day, MDS job - 3 days.
<table>
<thead>
<tr>
<th>COG</th>
<th>Annotation</th>
<th>Uniref100</th>
</tr>
</thead>
<tbody>
<tr>
<td>COG1131</td>
<td>ABC-type multidrug transport system, ATPase component</td>
<td>14406</td>
</tr>
<tr>
<td>COG1136</td>
<td>ABC-type antimicrobial peptide transport system, ATPase component</td>
<td>7306</td>
</tr>
<tr>
<td>COG1126</td>
<td>ABC-type polar amino acid transport system, ATPase component</td>
<td>4061</td>
</tr>
<tr>
<td>COG3839</td>
<td>ABC-type sugar transport systems, ATPase component</td>
<td>4121</td>
</tr>
<tr>
<td>COG0444</td>
<td>ABC-type dipeptide/oligopeptide/nickel transport system ATPase comp</td>
<td>3520</td>
</tr>
<tr>
<td>COG4608</td>
<td>ABC-type oligopeptide transport system, ATPase component</td>
<td>3074</td>
</tr>
<tr>
<td>COG3842</td>
<td>ABC-type spermidine/putrescine transport systems, ATPase comp</td>
<td>3665</td>
</tr>
<tr>
<td>COG0333</td>
<td>Ribosomal protein L32</td>
<td>1148</td>
</tr>
<tr>
<td>COG0454</td>
<td>Histone acetyltransferase HPA2 and Related acetyltransferases</td>
<td>14085</td>
</tr>
<tr>
<td>COG0477</td>
<td>Permeases of the major facilitator superfamily</td>
<td>48590</td>
</tr>
<tr>
<td>COG1028</td>
<td>Dehydrogenases with different specificities</td>
<td>37461</td>
</tr>
</tbody>
</table>

Visualizing PSU

ECMLS 2012
Heatmap of NW vs Euclidean Distances
Dendrogram of Cluster Centroids

Visualizing PSU
Selected Clusters
Heatmap for Selected Clusters

Heat Map of Euclidean Vs Transformed NW

Histogram of Transformed Distances

Visualizing PSU
Future Steps

- Comparison Needleman-Wunsch v. Blast v. PSIBlast
  - NW easier as complete; Blast has missing distances
- Different Transformations distance \( \rightarrow \) monotonic function(distance) to reduce formal starting dimension (increase sigma/mean)
- Automate cluster consensus finding as sequence that minimizes maximum distance to other sequences
- Improve \( O(N^2) \) to \( O(N) \) complexity by interpolating new sequences to original set and only doing small regions with \( O(N^2) \)
  - Successful in metagenomics
  - Can use Oct-tree from 3D mapping or set of consensus vectors
  - Some clusters diffuse?
Full Data Blast\textsuperscript{6} Original run has 0.96 cut
Cluster Data Blast 6
Original run has 0.96 cut
Use Barnes Hut OctTree originally developed to make $O(N^2)$ astrophysics $O(N \log N)$
OctTree for 100K sample of Fungi

We use OctTree for logarithmic interpolation
440K Interpolated

haixu_440k_100k_level4_20-100 Interpolation
Conclusions

- Data → Knowledge: protein annotation
  - Overwhelming influx of new sequences
  - Annotation is an immense challenge.
  - HPC and advanced analytics needed.

- PSU as tool to facilitate annotation:
  - Interactive visualization and exploration
  - Integrates info on function, pathways, structure, and environment
  - MDS preserves grouping structure of protein space
  - MDS can use different proximities and biological data
  - Parallel MDS handles large-scale data
  - MDS interpolation quickly maps new sequences into existing space
Data-Enabled Life Sciences Alliance International

- Collective innovation to tackle modern biological challenges through best computational practices and advanced cyberinfrastructure.
- Harness expertise and resources across disciplines
- Promote accurate, sustainable, scalable approaches
- Facilitate translation of data influx into tangible innovations and groundbreaking discoveries
References and Resources

- COG data is available at the NCBI site

- MDS results are available at
  http://manxcatcogblog.blogspot.com/

- All software used to analyze and visualize the data is an open source.

- DELSA: http://www.delsaglobal.org
  - Protein Global Atlas and Data Accessibility Projects
Grant support

- NSF: under DBI: 0969929 (EK) and 0910818 (GF)
- NIH: 5 RC2 HG 005806-02 (GF); NIGMS grant R01 GM-076680-04 (EK); NIDDK grants U01-DK-089571 and U01-DK-072473 (EK)
Thank you for your attention