High Performance Computing Workflow for Protein Functional Annotation

Larissa Stanberry\textsuperscript{1,2}
larissa.stanberry@gmail.com

Bhanu Rekepalli\textsuperscript{2,3}
brekapal@utk.edu

\textsuperscript{1}Seattle Children’s Research Institute, \textsuperscript{2}Data-Enabled Life Sciences Alliance, \textsuperscript{3}Joint Institute for Computational Sciences, University of Tennessee – Oak Ridge National Laboratory

Jointly with P. Giblock\textsuperscript{3}, Y. Liu\textsuperscript{3}, R.Higdon\textsuperscript{1,2}, W.Broomall\textsuperscript{1}, N.Kolker\textsuperscript{1,2} and E.Kolker\textsuperscript{1,2}

24 July 2013
Overview

- Grand challenge in functional genomics.
- Our goal.
- Solution, implementation, results.
- Community efforts.
Grand Challenge of Functional Genomics

- Functional annotation of protein sequences.
- Data space expands exponentially
  - EMP, i5K, iPlant, NEON.
  - 30% of sequenced proteins unknown function.
- Lack of tools, resources, cyberinfrastructure.
- Sustainability issues: databases, tools, CI.
- Annotation projects overwhelmed, many unsupported:
  - COG, Systers, ClusTr.
Our goal: HPC Annotation Pipeline

- Create automated pipeline for functional annotation.
- Use existing parallel bioinformatics applications on High Performance Computing resources.
- Develop sustainable software framework
- Make pipeline available to scientific community.
- Provide rigorous, reliable annotation tools.
Revitalize, expand & enhance bacterial COGs

- Clusters of Orthologous Groups of Proteins
- Developed by NCBI.
- Groups of proteins with common functions.
- Prokaryotes (COG): 66 genomes, 200K proteins, 5K clusters.
- Eukaryotes (KOG): 7 genomes, 113K proteins, 5K clusters.
- Valuable scientific resource: 5K citations.
- Newly sequenced bacterial genomes data: over 7.6 million sequences (version 22 April 2013).
Classification Algorithm

- COG
  - Train
  - Test
  - Compute Profiles
    - Psi-Blast
      - ID: Protein – top hit COG
        - If score > thresh: Protein → top hit COG
          - Validation: ROC

wc-COGs: 144K in 2,824 COGs
pc-COGs: 37K in 2,049 COGs
Compute tools
Algorithm Validation

- Archaeal proteins:
  - 120 archaeal genomes classified into COGs.
- Use COG profiles to calculate PSI-BLAST scores.
- Estimate true and false positive rates.
Performance: ROC curves

Figure: (left) Test data: FPR=10.0% and TPR=88.1%; (right) Archaeal proteins: FPR=21.9% and TPR=87.9% at log-threshold=4.0 (red).
Processing Scheme

0. Given: COG database
   New bacterial genomes DB

1. Compute COG profiles & consensuses

2. Select 200,000 proteins from new bacterial genomes

3. Compute PSI-BLAST scores for each protein vs COG consensus & profile

4. ID top hit COG for each protein. If PSI-BLAST score > threshold: Protein → top hit COG

5. With expanded COGs, re-evaluate threshold via test/train approach.

EXIT

Is new DB exhausted?

YES

NO
Tools and Parallel Implementations

- BLAST functions, MUSCLE, Perl and R scripts.
- Parallel implementations achieve scalability through database fragmentation, parallel I/O, load balancing, and query prefetching.
  - mpiBLAST, ScalaBLAST, ClustalW-MPI
- Highly scalable implementations:
  - BLAST on Blue Gene/L, pioBLAST, HSPp-BLAST.
- HSPp-BLAST validated on Kraken.
The Wrapper Architecture

- **Main components:** `mcw`, `stdiowrap`.
- `mcw` MPI app (C), hierarchy: master, controller, worker.
- `stdiowrap` provides I/O redirection through IPC implemented via shared memory segments.
Optimizations

- Each tool instance requires a copy of the database from the distributed file system
  - Performs poorly if thousands of processes read the same files simultaneously.
  - Master node reads database and broadcasts with MPI_Bcast.
  - Scales logarithmically with the number of nodes.

- To reduce input latency, sequences are read by master and prefetched by controllers.

- Two-stage buffering provides asynchronous writes.
  - Tools write to in-memory buffer.
  - Data is flushed (and optionally compressed) to disk when buffers are nearly full.
  - Increases output bandwidth, results in more uniform output time, and almost eliminates blocking.
Results Summary

- Processed one million proteins in five iterations.
- Mean (± SD) output size 31 ± 2.8 Gigabytes.
- 635,340 proteins assigned to COGs.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>140,375</td>
<td>4,408</td>
<td>5.2</td>
<td>86</td>
</tr>
<tr>
<td>2</td>
<td>126,813</td>
<td>4,251</td>
<td>5.4</td>
<td>87</td>
</tr>
<tr>
<td>3</td>
<td>123,196</td>
<td>4,152</td>
<td>5.4</td>
<td>89</td>
</tr>
<tr>
<td>4</td>
<td>123,411</td>
<td>4,175</td>
<td>5.4</td>
<td>90</td>
</tr>
<tr>
<td>5</td>
<td>121,545</td>
<td>4,139</td>
<td>5.4</td>
<td>90</td>
</tr>
</tbody>
</table>
Profile Compute Time

- Newton HPC (UT): 250 nodes with 32GB system memory, 2400+ cores (Intel Xeon E5-2670, 2.6 GHz).
- Compute time varied wrt cluster size and heterogeneity.
- COGs with <500 sequences: $O(\sqrt{n})$.
- COGs with >500 sequences: $O(n^2)$. 
RHS: COGs with >1,000 sequences after iteration 5.

- COG0583 (black) the largest: 6,231 sequences.

- K -transcription, T - signal transduction mechanisms.
Future Work

- Implementation is not fully automated.
- Perl scripts will be implemented in C for speed and compatibility.
- Extend the wrapper to run tools in a “streaming” fashion: data from step 1 is available at step 2 via shared memory segments.
  - Streaming avoids the cost of writing and reading temp files; provides increased data locality.
  - Compare to course-grained approach (Airavata) where each step is performed as a separate job.
Promise and Challenge of Big Data

- **Data ⇒ Knowledge**
  - Overwhelming influx of data.
  - Analysis and annotation an immense compute challenge.
  - Cross-disciplinary skills and expertise.
  - HPC and advanced analytic tools are needed.
  - Proposed: HPC workflow and low-complexity classification for annotation.

- **Knowledge ⇒ Action**
  - Advance knowledge of biological systems.
  - Understand, predict, diagnose diseases
  - Discover biomarkers, drug targets.
  - Generate hypothesis, design experiments.
What Future Holds

- **OLD**: One PI - Jack of all trades.
  - Instrument support, data acquisition, pre-processing, analysis, software, compute resources, IT support, publications.
- **NEW**: Community innovation & x-disciplinary efforts
  - Join forces to meet challenges of data-enabled sciences.
  - Share data, skills, tools.
  - Provide robust, sustainable, scalable resources.
  - Use best compute and analytic practices.
Data-Enable Life Sciences Alliance Global

- Trans-disciplinary alliance of scientists and experts.
- Promotes collective innovation through interdisciplinary research.
- Translate influx of data into tangible innovations.
Acknowledgements

- Eugene Koonin and Yuri Wolf (NCBI) for expert advice and technical help.
- Feedback and comments from reviewers.
- NSF 0969929, NIDDK U01-DK-089571 and U01-DK-072473, SCRI, The Robert B. McMillen Foundation award, and The Gordon and Betty Moore Foundation award to EK
- XSEDE allocation of advanced computing resources (Kraken Supercomputer) was provided by the NSF and NSF grant EPS-0919436 to BR.
Thank you for your attention.