Problems, Challenges and Opportunities in Exploring the “Dark Matter” of Life Sciences: The Microbiome

Ananth Kalyanaraman
Boeing Centennial Chair in CS
Associate Professor
School of EECS
Washington State University
ananth@eecs.wsu.edu
http://www.eecs.wsu.edu/~ananth

ParLearning 2015
Held in conjunction with IPDPS
Hyderabad, India, May 29, 2015
Talk Outline

(\textit{the area})

- An Overview on Microbiome
  (\textit{what we have been doing?})

- Scalable graph-theoretic models for functional characterization of microbiomes
  (\textit{how you can contribute?})

- Future trends, challenges and opportunities
Microbiome: Understanding the Microbial Universe

- **Number of microbial species:** unknown
- **Estimates:** unknown (<1% known?)

> “What we know is the size of a fistful of soil, what we don’t know is the size of this world.”

“காண்பது எத்தனையுமே, கூற்றல்லது என்றழையுமே”

Avvaiyar, 1st century CE

- **Microbes are the Engineers of Nature**
Microbiomes of our Environment with Applications

Extremophiles

Medicine

Carbon Sequestration

Env. bioremediation

Ag. biotech.

Biofuels

Green house gas reduction
“Metagenomics”

- Application of genomics tools to study microbial communities in their natural environment
  - Without necessitating the lab cultivation and culture of individual genomes
Metagenomics: A Big Data challenge

**Volume**
- $10^9$ sequences from a single experiment
- 1 Gbp per gram of soil
- thousands of species

**Variety**
- species diversity
- metagenomic DNA
- metaproteomics
- RNAseq
- microarrays
- pathways
- ...

**Velocity**

565 projects as of May 2015

![Graph showing project growth from 1997 to 2011]

Term coined

**Funding**

![Pie chart showing project funding]

Charts from: http://www.genomesonline.org
Who is out there?

What are they doing?

FIG. 3. Construction and screening of metagenomic libraries. Schematic representation of construction of libraries from environmental samples. The images at the top from left to right show bacterial mats at Yellowstone, soil from a boreal forest in Alaska, cabbage white butterfly larvae, and a tube worm.
How to Assemble a Genome?

Input: Multiple copies of the genome

Randomly cut each copy and generate copies

Output: Unordered genome fragments

Next Task: Assemble the genome from its fragments

Randomly cut each copy and generate copies
Like a Jigsaw Puzzle (not a perfect analogy but close enough)

Primary source of information for assembly:
Overlapping reads

Overlaps should account for errors and mutations

Assembling metagenomes $\equiv$ assembling multiple jigsaw puzzles
Application Framework for Functional Characterization

Obtain samples

Generate DNA reads ($10^7$-$10^{10}$)

Assemble metagenome & predict genes

Microbial protein databases

~$10^6$ clusters

$10^8$ sequences

Sequence analysis & graph-theoretic modeling and analysis

- Translate into
- Open Reading Frames (ORFs, $10^6$-$10^8$)

...
Establishing pairwise relationship

Q) How to tell whether two proteins are related?

MTVKEQSDIVHGIMSQCVLM

MTSKEQS - -VHDIMSQCVLM

Sequence-level similarity

MTSKEQSVHDIMSQCVLM

Structural similarity

Other models:
• Observed interactions
• Transcriptomics
• Predict interactions
The protein clustering pipeline
(Yooseph et al. 2007, Wu & Kalyanaraman, 2008)

$n$ sequences

All vs. All sequence comparison

Remove redundant sequences

G(V,E)

Homology Graph

Protein clusters

Community annotation
Related work & challenges

- **Homology detection**
  - BLAST based (e.g., Yooseph et al. 2007)
  - Low sensitivity due to the heuristic nature
    - Between 40% - 70% edges missed (Wu et al. 2012)
  - Smith-Waterman based methods ruled out presently in practice!

- **Clustering**
  - NP-hard formulations
  - In practice, heuristics with no qualitative guarantees are used (e.g., CD-HIT, UCLUST)
  - Lack of scalable tools (DIMACS’10 challenge)
Our contributions

1) **Scalable homology graph construction**
   - Optimal alignment computation (Smith-Waterman) at scale
   - Uses suffix trees/ESA for exact match-based filtering
   - Parallel tool – scales to 1K+ cores;
     \[ \times 10^6 \text{ sequence inputs in minutes} \]
   - Selected Papers: Wu et al. TPDS 2012, Daily et al. JPDC 2014

2) **Scalable graph heuristics for clustering**
   - Two different heuristics:
     i) Modularity detection; and ii) Dense subgraph detection
   - OpenMP and MapReduce-MPI implementations
   - Qualitative improvements over other heuristics
   - Scales to tens of cores; \( \times 10^6 \) sequence inputs in minutes
   - Selected Papers:
     Dense subgraphs: Wu et al. SC|08, Rytsareva et al. ICCABS 2013
Joint work with: Jeff Daily, Sriram Krishnamoorthy @ PNNL (JPDC 2014, HiPC 2012)

PGRAPH-TASCEL:
HOMOLOGY DETECTION
Exact matching filter: Suffix Trees

pairwise alignments

s_1

s_2

x x x

m

n
Exact matching filter: Suffix Trees

generalized suffix tree filter

0: ABAB$_0$
1: BABA$_1$
2: ABBA$_2$

pairwise alignments

Data dependencies

image source: Wikipedia, Nils Grimsmo

Parallel Generation of Suffix Trees

Number of buckets: $|\Sigma|^k$ where $k$ is window length. e.g., $k=2$

Generate all pairs with match length $\geq M$, e.g., proteins $M=7$
Load balancing strategies

Dynamic task counters

- Counter on a single rank
- (Remote) atomic read and increment

Message Bus

P1
P2
P3
P4

But we don’t know the tasks ahead of time!
Work Stealing for Homology Detection

Input sequences on a shared buffer

\[ s_1 \quad s_2 \quad \ldots \quad s_n \]

Worker threads

Deques

Helper thread

Key

- Orange: Private portion
- Gray: Shared portion
- Task delivery
- Task theft

Mixed task types

- T=Tree
- SW=Alignment

Incoming/outgoing steal requests

Network interconnect
pGraph-Tascel: Experimental Results

**Time to solution (vs. p)**

![Graph showing time to solution vs. p for different sequence sizes on different numbers of cores.](image)

**Hopper (NERSC-6): Cray XE6**
- 6,392 compute nodes, 153,216 cores
- 32 GB RAM per node
- Custom mpich-2 version 5.5.5 for Cray XE
Joint work with: Hao Lu, Mahantesh Halappanavar, Daniel Chavarria
(PARCO 2015, HiPC 2014, IPDPS 2015)

GRAPPOLO: MODULARITY-BASED CLUSTERING HEURISTICS
Graph clustering

- **Problem:** Given $G(V, E, \omega)$, identify tightly knit groups of vertices that strongly correlate to one another within their group, and sparsely so, outside.

**Input:**
- $V = \{1, 2, \ldots, n\}$
- $E$: a set of $M$ edges
- $\omega(e)$: weight of edge $e$ (non-negative)
- $m = \sum_{e \in E} \omega(e)$

**Output:**
- A partitioning of $V$ into $k$ mutually disjoint clusters $P = \{C_1, C_2, \ldots, C_k\}$ such that: …?
Modularity (Newman 2004)

- A statistical measure for assessing the quality of a given community-wise partitioning $P$ of the vertices $V$:

$$Q = \frac{1}{2m} \sum_{i \in V} e_{i \rightarrow C(i)} - \sum_{C \in P} \left( \frac{a_C}{2m} \cdot \frac{a_C}{2m} \right)$$

Fraction of intra-cluster edges

Equivalent fraction in a random graph
Louvain method (Blondel et al. 2008)

Input: G(V,E)
Goal: Compute a partitioning of V that maximizes modularity (Q)
Init: Every vertex starts in its own community (i.e., $C(i) = \{i\}$)

Multi-phase iterative heuristic
Within each iteration:

- **For** every vertex $i \in V$:
  1. Let $C(i)$: current community of $i$
  2. Compute modularity gain ($\Delta Q$) for moving $i$ into each of $i$'s neighboring communities
  3. Let $C_{max}$: neighboring community with largest $\Delta Q$
  4. If ($\Delta Q > 0$) { Set $C(i) = C_{max}$ }
Louvain method (Blondel et al. 2008)

Input: G(V,E)
Goal: Compute a partitioning of V that maximizes modularity (Q)
Init: Every vertex starts in its own community (i.e., $C(i)=\{i\}$)

Multi-phase iterative heuristic
Within each iteration:

- **For** every vertex $i \in V$:
  1. Let $C(i)$: current community of $i$
  2. Compute modularity gain ($\Delta Q$) for moving $i$ into each of $i$'s neighboring communities
  3. Let $C_{\text{max}}$: neighboring community with largest $\Delta Q$
  4. If ($\Delta Q>0$) { Set $C(i) = C_{\text{max}}$ }
Our Approach: **Grappolo** (Lu et al. Parallel Computing 2015)

A parallel implementation for the Louvain heuristic

- Multiple graph heuristics for parallelization
- Graph coloring to improve results and balanced coloring to improve load balancing (Lu et al. IPDPS 2015)
- Acceleration on the Tilera many-core architecture (Chavarria et al., HiPC 2014)
- Linear scaling achieved with relative speedups up to 47x on 36 cores over serial

http://hpc.pnl.gov/people/hala/grappolo.html
Parallelizing the Louvain heuristic

- Simple parallelization could delay convergence or cause negative modularity gain
  - Use 1-distance coloring

- To prevent vertex swaps use min/max labeling

- Detect phase transitions to get over local maxima

- Preprocess single-degree vertices (“vertex following”)

(Lu et al. PARCO’15)
Performance results

Relative speedup

Convergence rate and modularity
Comparison with serial Louvain

<table>
<thead>
<tr>
<th>Input</th>
<th>Output modularity</th>
<th>Run-time (in sec)</th>
<th>Speedup (8 threads)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Parallel</td>
<td>Serial (8 threads)</td>
<td></td>
</tr>
<tr>
<td>CNR</td>
<td>0.912608</td>
<td>0.912784</td>
<td>5.37×</td>
</tr>
<tr>
<td>coPapersDBLP</td>
<td>0.858088</td>
<td>0.848702</td>
<td>2.08×</td>
</tr>
<tr>
<td>Channel</td>
<td>0.933388</td>
<td>0.849672</td>
<td>1.45×</td>
</tr>
<tr>
<td>Europe-osm</td>
<td>0.994996</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>MG1</td>
<td>0.968723</td>
<td>0.968671</td>
<td>4.39×</td>
</tr>
<tr>
<td>uk-2002</td>
<td>0.989569</td>
<td>0.9897</td>
<td>1.59×</td>
</tr>
<tr>
<td>MG2</td>
<td>0.998397</td>
<td>0.998426</td>
<td>2.86×</td>
</tr>
<tr>
<td>NLPKKT240</td>
<td>0.934717</td>
<td>0.952104</td>
<td>13.07×</td>
</tr>
<tr>
<td>Rgg_n_2_24_s0</td>
<td>0.992698</td>
<td>0.989637</td>
<td>3.24×</td>
</tr>
<tr>
<td>Soc-LiveJournal</td>
<td>0.751404</td>
<td>0.726785</td>
<td>2.72×</td>
</tr>
<tr>
<td>friendster</td>
<td>0.626139</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>
THE CARBON CYCLE

Crops like corn are finely ground and separated into their component sugars.

that is reabsorbed by the original crops.

CO₂

which releases carbon dioxide which can be used as an alternative fuel

The sugars are distilled to make ethanol,

A CASE STUDY: THE BIOFUELS CHALLENGE

Image credit: http://sitemaker.umich.edu/
FOOD OR FUEL?

Nearly a billion people will go hungry tonight, yet this year the U.S. will turn nearly 5 billion bushels of corn into ethanol. That’s enough food to feed 412 million people for an entire year.

8 BUSHELS OF CORN = 21.6 GALLONS OF ETHANOL FUEL OR ENOUGH FOOD TO FEED A PERSON FOR A WHOLE YEAR

http://foodorfuel.weebly.com/

EXPLAIN THE MATH...

5 billion bushels / 8 bushels of corn (enough caloric feed for a person for a year) = sufficient calories to support 625 million people, minus one third to account for distiller's grain (80% = 42% maize)

8 bushels of corn feeds a person for a year

X 2.7 gallons of ethanol per bushel = 21.6 gallons of ethanol per bushel

SOURCES

400 pounds of corn supplies enough calories for one person for a year (http://www.forkastnews.com/articles/asb100-feeding-a-person-how-biofuels-could-starve-the-planet/)

About 5 billion bushels of U.S. corn production is deleted for ethanol production (http://www.csdl.siu.edu/commoditynews/illiant.pdf)

One bushel of corn produces 2.7 gallons of ethanol (Purdue Extension, "Corn: Fuel Ethanol is Made From Corn," http://www.extension.purdue.edu/extmedia/ID-328.pdf)
Wht are the alternatives?

Corn

Switchgrass

Image credit: Jonas Lovaas Gjerstad
The Cow Rumen Solution

(Hess et al. Science 2011)
The Cow Rumen Project  
(Hess et al. Science 2011)

- Key findings:
  - Switchgrass inserted through the cannula gets degraded in 72 hours
  - The cow gut has microbes who do the job
  - Sequencing of the cow gut “microbiome” leads to:
    - 268GB of DNA sequenced
    - 27,755 candidate genes

- Next step:
  - Proteins and Pathways

Image credit: http://preventcancer.aicr.org/
Cow Rumen Data Analysis

Goal: to understand the microbial machinery behind cellulose degradation in cow rumen

Detect homology using pGraph (Daily et al. 2014)

- 2,547,270 sequences
- 95,701,413 edges

Clustering statistics:

<table>
<thead>
<tr>
<th></th>
<th>Grappolo</th>
</tr>
</thead>
<tbody>
<tr>
<td># seq. in clusters</td>
<td>2,361,667 (92.7%)</td>
</tr>
<tr>
<td># clusters</td>
<td>206,813</td>
</tr>
<tr>
<td>Largest cluster</td>
<td>23,611</td>
</tr>
<tr>
<td>Modularity</td>
<td>0.88</td>
</tr>
<tr>
<td>Density</td>
<td>0.79 +/- 0.23</td>
</tr>
<tr>
<td>Time</td>
<td>17.80 sec (32 threads, 5.6x speedup)</td>
</tr>
</tbody>
</table>
Cow Rumen Data Analysis

Goal: to understand the microbial machinery behind cellulose degradation in cow rumen

(45 mins, 2K cores)

Detect homology using pGraph (Daily et al. 2014)

2,547,270 sequences → 95,701,413 edges

Clustering statistics

<table>
<thead>
<tr>
<th></th>
<th>Grappolo</th>
<th>BLAST-graph</th>
</tr>
</thead>
<tbody>
<tr>
<td># seq. in clusters</td>
<td>2,361,667 (92.7%)</td>
<td>783,948 (30.8%)</td>
</tr>
<tr>
<td># clusters</td>
<td>206,813</td>
<td>10,106</td>
</tr>
<tr>
<td>Largest cluster</td>
<td>23,611</td>
<td>4,829</td>
</tr>
<tr>
<td>Modularity</td>
<td>0.88</td>
<td>0.92</td>
</tr>
<tr>
<td>Density</td>
<td>0.79 +/- 0.23</td>
<td>0.33 +/- 0.30</td>
</tr>
<tr>
<td>Time</td>
<td>17.80 sec (32 threads, 5.6x speedup)</td>
<td>9 sec</td>
</tr>
</tbody>
</table>
Protein family characterization: Progress Report

1. Picked representative sequences from 206,813 clusters

2. Performed a PFAM search for those clusters, and extracted 911 clusters that mapped to cellulose degradation domains

3. Performed a \texttt{tblastn} of these 911 representatives vs. 2.5M predicted genes
   1. 911 reps mapped to 906 distinct genes

4. Genes are being sequenced using PCR
An ongoing application – Ocean data (collaboration with Hallam lab, UBC)

Goal: to understand the microbial mechanisms under oxygen-deficient settings

(15 mins, 32K cores)

Detect homology using pGraph (Daily et al. 2014)

11,005,829 seqs. → 685,250,898 edges

Clustering statistics

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td># sequences in clusters</td>
<td>10,346,702 (94.01%)</td>
</tr>
<tr>
<td># clusters</td>
<td>1,346,405</td>
</tr>
<tr>
<td>Largest cluster</td>
<td>14K</td>
</tr>
<tr>
<td>Modularity</td>
<td>0.9856</td>
</tr>
<tr>
<td>Time</td>
<td>214 sec (32 threads, 5.1x speedup)</td>
</tr>
</tbody>
</table>

Degree distribution obeys Power Law ($\lambda = -2.487$)

Cluster size distribution
ALGORITHMIC CHALLENGES, FUTURE TRENDS, AND OPPORTUNITIES
Challenges in Sequencing a Microbial Community

- **“Culture Divide”:**
  - Less than 1% microbial genomes can be cultured individually in a laboratory!

- **Diversity:**
  - Communities are comprised of thousands of species

- **Non-uniformity:**
  - High variation in species abundance
Identification of Species: Taxonomical binning

- A problem of “census”
  - A fundamental descriptor of the community structure in terms of its species composition
- Most popular approach:
  - “Species” ≡ “Ribotypes”: Identify conserved phylogenetic markers (16S rRNA)

Image credit: Alice McHardy
Current Approaches for Taxonomical Binning

Marker Gene Analysis → Marker Gene Database Comparison → Taxonomic/Phylogenetic Classification → Taxon or Phylogenetic Abundance Profile

Composition Classification/Clustering → Classification/Clustering Algorithm → Taxon Abundance Profile

Sequence Database Comparison → Taxonomic/Phylogenetic Classification → Taxon or Phylogenetic Abundance Profile

Genome Sequence Database Comparison → Fragment Recruitment → Genome or Contig Coverage Profile

Assembly → Contig Assembly → Supercontig/Genome Assembly → Genome or Contig Coverage Profile

Credit: Sharpton, 2014
Identification of Species…

- Perhaps the most reliable source of census should be the assemblers
- Sadly, current assemblers still at large rely on the ribotypes from PCR experiments

- **How to design an effective, inbuilt, on-the-fly species differentiator within assemblers?**
- Potential directions:
  - Consider genomic “signatures” (eg., codon bias, GC content, CpG frequency)

![Self-Organizing Maps of tetratnucleotide usage for fragments of 81 prokaryotic genomes (coloring by species). From: Abe et al. DNA Research 2005](image1)

![3-D display of 7-letter frequencies for 100 kb segments of 4 organisms. From: Deschavanne et al., Mol. Biol. Evol. 1999](image2)
Identification of Species…

- Perhaps the most reliable source of census should be the assemblers
- Sadly, current assemblers still at large rely on the ribotypes from PCR experiments

**How to design an effective, inbuilt, on-the-fly species differentiator within assemblers?**

- Potential directions:
  - Consider genomic “signatures” (e.g., codon bias, GC content, CpG frequency)
  - Incorporate analytical capabilities from comparative genetics and population genetics
  - Provide phylogenetic input during assembly
  - Shift from a “genome-centric” (structure) to “gene-centric” approach (function)
Validation of Assemblies

- Large fractions of unassembled data
  - E.g., 53% of reads remained singletons in the Sorcerer II project
- Fragmented assemblies with much smaller contigs
  - Only 9% of reads assembled into contigs of length > 10Kbp
- Differentiate between repeats vs. genuine duplicate copies
  - Need a method to weed out redundancy but also learn from it

**Conventional strategy**

Rep #1 — Rep #2 — Rep #3

Too many overlaps.
So discard as repeat region

**Metagenomics scenario**

Too many overlaps but legitimate.
So do not discard!
Comparative Metagenomics

- How to compare communities at different scales and obtained from different time steps?

Source: Kelly et al., PNAS 2003
Knowledge Graphs for Microbiomes?

A heterogeneous network for metagenomics/metaproteomics:

Integration of heterogeneous information can provide new levels of insights into the functioning of a naturally built system.
The Future: Longitudinal Surveys

- Deep sequencing of individual communities

- How do the communities adapt to changing environment?

- Implications:
  - Gastrointestinal disorders
  - Anti-biotic microbial resistance
  - Personalized medicine
  - Agricultural biotechnology
The Future: Longitudinal Surveys

- Microbiome survey of Families vs. Households (Lax et al. 2014)
The Future: Longitudinal Surveys

Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization (Sharon et al. 2013)

- 260M reads, 11 time points (2.4Gbp per sample)
Concluding Remarks

- Microbial community sequencing and their functional analysis projects have phenomenal practical implications.

- Advancements in the following aspects are key for all future endeavors in metagenomics:
  - Sequencing technologies & strategies
  - Better analytical methods and new functions
  - Role of HPC cannot be understated but needs to happen with algorithmic innovations and new ways to model

- Data to Models?
DATA + SCIENCE = DATA SCIENCE

Raw data (next-gen & traditional)

Finished sequences (genomes, genes, proteins, CDS)

Annotation & metadata
Acknowledgements

Students contributors:
- Hao Lu, Jeff Daily, Inna Rytsareva, Andy Wu

Collaborators:
- Mahantesh Halappanavar, Daniel Chavarria @ PNNL
- Assefaw Gebremedhin @ WSU
- Matthias Hess @ WSU/UC Davis
- Steven Hallam @ Univ. British Columbia
- Sriram Krishnamoorthy @ PNNL

Support:
- DOE DE-SC-0006516, NSF IIS 0916463, NSF DBI 1262664
- NERSC