Streaming and Compression Approaches for Terascale Biological Sequence Data Analysis

C. Titus Brown
Assistant Professor
CSE, MMG, BEACON
Michigan State University
September 2012
ctb@msu.edu
Outline

• Acknowledgements

• Big Data and next-gen sequence analysis

• Sweeping generalizations about physics and biology

• Physics ain’t biology, and vice versa
Acknowledgements

Lab members involved

- Adina Howe (w/Tiedje)
- Jason Pell
- Arend Hintze
- Rosangela Canino-Koning
- Qingpeng Zhang
- Elijah Lowe
- Likit Preeyanon
- Jiarong Guo
- Tim Brom
- Kanchan Pavangadkar
- Eric McDonald

Collaborators

- Jim Tiedje, MSU
- Janet Jansson, LBNL
- Susannah Tringe, JGI

Funding

USDA NIFA; NSF IOS; BEACON.
We practice open science!

See blog post accompanying talk: ‘titus brown blog’

Everything discussed here:

- Code: github.com/ged-lab/ ; BSD license
- Blog: http://ivory.idyll.org/blog
- Twitter: @ctitusbrown
- Grants on Lab Web site: http://ged.msu.edu/interests.html
- Preprints: on arXiv, q-bio:
  ‘diginorm arxiv’
Soil is full of uncultured microbes

Randy Jackson
Soil contains thousands to millions of species
(“Collector’s curves” of ~species)

99% of microbes cannot easily be cultured in the lab.
Shotgun metagenomics

- Collect samples;
- Extract DNA;
- Feed into sequencer;
- Computationally analyze.

Wikipedia: Environmental shotgun sequencing.png
Task: assemble original text from random, error prone observations

It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness.
Actual coverage varies widely from the average.
Assembly via de Bruijn graphs – k-mer overlaps

(a) aaccgg
    ccggtt

(b) aacc → accg → cçgg → cggt → ggtt

(c) aaccgggtt
K-mer graph (k=14)

Single nucleotide variations cause long branches;
They don’t rejoin quickly.
Reads vs edges (memory) in de Bruijn graphs

#Edges

#Reads

Total edges

Error edges

True edges

Conway T C, Bromage A J Bioinformatics 2011;27:479-486
The scale of the problem is stunning.

- I estimate a worldwide capacity for DNA sequencing of 15 petabases/yr (it’s probably larger).
- Individual labs can generate ~100 Gbp in ~1 week for $10k.
- This sequencing is at a boutique level:
  - Sequencing formats are semi-standard.
  - Basic analysis approaches are ~80% cookbook.
  - Every biological prep, problem, and analysis is different.

- Traditionally, biologists receive no training in computation. (And computational people receive no training in biology :)

- …and our computational infrastructure is optimizing for high performance computing, not high throughput.
My problems are also very annoying...

- Est ~50 Tbp to comprehensively sample the microbial composition of a gram of soil.
- Currently we have approximately 2 Tbp spread across 9 soil samples, for one project; 1 Tbp across 10 samples for another.

- Need 3 TB RAM on single chassis to do assembly of 300 Gbp.
- …estimate 500 TB RAM for 50 Tbp of sequence.

That just won’t do.
1. Compressible de Bruijn graphs

Each node represents a 14-mer;
Links between each node are 13-mer overlaps
Can store *implicit* de Bruijn graphs in a Bloom filter

This allows *compression* of graphs at the expense of false positive nodes/edges.
False positives introduce false nodes/edges.

When does this start to distort the graph?
Global graph structure is retained past 18% FPR
Equivalent to bond percolation problem; percolation threshold independent of $k$ (?)
This data structure is strikingly efficient for storing sparse k-mer graphs. “Exact” is for best possible information-theoretical storage.

Jason Pell & Arend Hintze
We implemented graph partitioning on top of this probabilistic de Bruijn graph.

Split reads into “bins” belonging to different source species.
Can do this based almost entirely on connectivity of sequences.
2. Online, streaming, lossy compression.

Much of next-gen sequencing is redundant.
Uneven coverage => even more redundancy

Suppose you have a dilution factor of A (10) to B(1). To get 10x of B you need to get 100x of A! Overkill!!

This 100x will consume disk space and, because of errors, memory.
Downsample based on de Bruijn graph structure; this can be derived via an online algorithm.
for read in dataset:
    if estimated_coverage(read) < CUTOFF:
        update_kmer_counts(read)
        save(read)
    else:
        # discard read

Note, single pass; fixed memory.
Digital normalization retains information, while discarding data and errors

Table 1. Digital normalization to C=20 removes many erroneous k-mers from sequencing data sets. Numbers in parentheses indicate number of true k-mers lost at each step, based on reference.

<table>
<thead>
<tr>
<th>Data set</th>
<th>True 20-mers</th>
<th>20-mers in reads</th>
<th>20-mers at C=20</th>
<th>% reads kept</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simulated genome</td>
<td>399,981</td>
<td>8,162,813</td>
<td>3,052,007 (-2)</td>
<td>19%</td>
</tr>
<tr>
<td>Simulated mRNAseq</td>
<td>48,100</td>
<td>2,466,638 (-88)</td>
<td>1,087,916 (-9)</td>
<td>4.1%</td>
</tr>
<tr>
<td><em>E. coli</em> genome</td>
<td>4,542,150</td>
<td>175,627,381 (-152)</td>
<td>90,844,428 (-5)</td>
<td>11%</td>
</tr>
<tr>
<td>Yeast mRNAseq</td>
<td>10,631,882</td>
<td>224,847,659 (-683)</td>
<td>10,625,416 (-6,469)</td>
<td>9.3%</td>
</tr>
<tr>
<td>Mouse mRNAseq</td>
<td>43,830,642</td>
<td>709,662,624 (-23,196)</td>
<td>43,820,319 (-13,400)</td>
<td>26.4%</td>
</tr>
</tbody>
</table>

Table 2. Three-pass digital normalization removes most erroneous k-mers. Numbers in parentheses indicate number of true k-mers lost at each step, based on known reference.

<table>
<thead>
<tr>
<th>Data set</th>
<th>True 20-mers</th>
<th>20-mers in reads</th>
<th>20-mers remaining</th>
<th>% reads kept</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simulated genome</td>
<td>399,981</td>
<td>8,162,813</td>
<td>453,588 (-4)</td>
<td>5%</td>
</tr>
<tr>
<td>Simulated mRNAseq</td>
<td>48,100</td>
<td>2,466,638 (-88)</td>
<td>182,855 (-351)</td>
<td>1.2%</td>
</tr>
<tr>
<td><em>E. coli</em> genome</td>
<td>4,542,150</td>
<td>175,627,381 (-152)</td>
<td>7,638,175 (-23)</td>
<td>2.1%</td>
</tr>
<tr>
<td>Yeast mRNAseq</td>
<td>10,631,882</td>
<td>224,847,659 (-683)</td>
<td>10,532,451 (-99,436)</td>
<td>2.1%</td>
</tr>
<tr>
<td>Mouse mRNAseq</td>
<td>43,830,642</td>
<td>709,662,624 (-23,196)</td>
<td>42,350,127 (-1,488,380)</td>
<td>7.1%</td>
</tr>
</tbody>
</table>
For soil... what do we assemble?

<table>
<thead>
<tr>
<th>Total Assembly</th>
<th>Total Contigs</th>
<th>% Reads Assembled</th>
<th>Predicted protein coding</th>
<th>rplb genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.5 bill</td>
<td>4.5 mill</td>
<td>19%</td>
<td>5.3 mill</td>
<td>391</td>
</tr>
<tr>
<td>3.5 bill</td>
<td>5.9 mill</td>
<td>22%</td>
<td>6.8 mill</td>
<td>466</td>
</tr>
</tbody>
</table>

This estimates number of species

Putting it in perspective:
Total equivalent of ~1200 bacterial genomes
Human genome ~3 billion bp

Adina Howe
Concluding thoughts

• Our approaches provide significant and substantial practical and theoretical leverage to one of the most challenging current problems in computational biology: assembly.

• They provide a path to the future:
  • Many-core compatible; distributable?
  • Decreased memory footprint => cloud computing can be used for many analyses.
  • At an algorithmic level, provide a noise-filtering solution for most of the current sequencing Big Data problems.

• They are in use, ~dozens of labs using digital normalization.

• …although we’re still in the process of publishing them.
Physics ain’t biology

The following observations are for discussion... they are not-so-casual observations from a lifetime of interacting with physicists.

(Apologies in advance for the sweeping generalizations.)
Important note: I don’t hate physicists!

Significant life events involving physicists:

Birth – Gerry Brown
First UNIX account – Mark Galassi
First publication – w/Chris Adami
Grad school plans – Hans Bethe et al.
Earthshine research (~8 pubs) – w/Steve Koonin and Phil Goode
2nd Favorite publication – w/Curtis Callan, Jr.

I am very physicist-positive!
1. Models play a very different role.

- Physics models are often *predictive* and *constraining*.
  - Model specifies dynamics or interaction.
  - Make specific measurements to obtain initial conditions.
  - Model can then be used to predict fine-grained outcomes.

- Biology models can rarely be built in the first place…
  - Models are dominated by unknowns.
  - In a few cases, can be used to determine *sufficiency* of knowledge ("the observations can be explained by our model"); this does not mean the model is correct, merely that it *could* be.
  - Models are rarely *predictive* of specific observations.
Endomesoderm network

Approximately 15 years and probably 200 man-years of research to assemble a map of gene interactions for the first 30 hours of sea urchin development.

http://sugp.caltech.edu/endomes/
Endomesoderm Specification up to 30 Hours

This model is frequently revised. It is based on the latest laboratory data, some of which is not yet published.


The current VFA includes not yet published co-regulatory data of Sinan de Leon, Joe Smith (in press), Andrew Cameron, Qiang Tu, Sagar Dharm, Andrew Luskin, Christina Theodoris, and, in addition to published data, is based on recent perturbation and other results of Isabelle Peter (endomesoderm domain), Stefan Mattern (NSM domain), and Joan Smith (NSP domain) of the Davidson lab. Relevant perturbation and expression data from these studies are presented here.

http://sugp.caltech.edu/endomes/
2. Little or no tradition of computation in biology

- Until ~last decade, not too much in the way of big data.
- Models are rarely built for the purpose of understanding computational data, although that is changing.
- Ecological and evolutionary models are regarded with suspicion: guilty until proven innocent.
- Essential zero computational training at UG/G (although some math).

- “Sick” culture of computation in biology:
  - Development of computational methods not respected as independent scientific endeavor in biology.
  - Biologists want push-button software that “just works”.
  - Sophisticated evaluation/validation of software by users is rare.

(It is hard for me to explain to biologists how big a problem this is.)
3. Biology is built on facts, not theory.

- Experience with Callan:
  - Constrained optimization of DNA binding model to 48 known CRP binding sites => inability to eliminate 300-3000 extra sites in *E. coli* genome.
  - Ohmigod their binding signature is preserved by evolution => they’re probably real! How can this be!?
  - …well, it turns out we don’t know that much about *E. coli*.

- A nice damning quote from Mark Galassi:
  “Biology and bioinformatics seem interesting. Is there any way I can take part in the research without learning all the details?”

  **NO. Biology is all about the details! The more the better!**
My career path

• Undergrad in Math (Reed)
  • Research on evolution model (Avida) ~1992
  • Earthshine observations of global albedo ~1994
• PhD in Molecular Developmental Biology (Caltech)
  • Molecular biology, genomics, gene regulation ~1997-2008
  • Bioinformatics ~2000-
• Faculty position in CSE and Microbiology (MSU), 2008
  • Molecular developmental biology
  • Bioinformatics
  • Metagenomics & next-gen sequence analysis more generally
  • Moving towards integration of data + modeling…
My career path

- Undergrad in Math (Reed)
  - Research on evolution model (Avida) ~1992
  - Earthshine observations of global albedo ~1994
- PhD in Molecular Developmental Biology (Caltech)
  - Molecular biology, genomics, gene regulation ~1997-2008
  - Bioinformatics ~2000-
- Faculty position in CSE and Microbiology (MSU), 2008
  - Molecular developmental biology
  - Bioinformatics
  - Metagenomics & next-gen sequence analysis more generally
  - Moving towards integration of data + modeling…
Concluding thoughts on this stuff

- Biologists simply don’t trust models and data analysis approaches that come from “outside” biology.
- They’re not necessarily wrong!
- Physicists can bring an important skill set and attitude to biological research, but their knowledge is useless. They have to meet the biology more than halfway.
- Biologists need more cross-training so we don’t retrace the same software development, data analysis, and modeling mistakes that physics et al. has spent 30 years figuring out.

But if you disagree with any of this, I’m happy to chat.