Algorithms for studying the structure and function of genomes

Michael Schatz
The double helix is a sheet of paper that genetic messages can be written upon.

The particular sequence of nucleotides in your genome, along with your environment and experiences, shapes who you are:

- Physical traits: Height, hair color, skin color, …
- Behavioral traits: Intelligence, Personality, …
- Susceptibility to disease, stress, and toxins
- Response to drug treatments

Finding changes to genome structure can provide powerful clues to its function.
Genomic Data

The instruments provide data, but none of the answers to any of our questions.

*Who will answer them?*

*How will they do it?*

Worldwide capacity exceeds 35 Pbp/year
Data Science Technologies

- Sensors & Metadata
  - Sequencers, Microscopy, Imaging, Mass spec, Metadata & Ontologies

- IO Systems
  - Hardrives, Networking, Databases, Compression, LIMS

- Compute Systems
  - CPU, GPU, Distributed, Clouds, Workflows

- Algorithmics
  - Streaming, Sampling, Indexing, Parallel

- Machine Learning
  - Classification, modeling, visualization & data integration

- Domain Analysis
System Level Advances

Optimizing data intensive GPGPU computations for DNA sequence alignment

CloudBurst: Highly Sensitive Read Mapping with MapReduce.

Design patterns for efficient graph algorithms in MapReduce.
Lin, J., Schatz, MC. (2010) Proceedings of the 8th Workshop on Mining and Learning with Graphs

The DNA Data Deluge
Data Science Technologies

- Sensors & Metadata
  - Sequencers, Microscopy, Imaging, Mass spec, Metadata & Ontologies

- IO Systems
  - Hardrives, Networking, Databases, Compression, LIMS

- Compute Systems
  - CPU, GPU, Distributed, Clouds, Workflows

- Algorithmics
  - Streaming, Sampling, Indexing, Parallel

- Machine Learning
  - Classification, modeling, visualization & data Integration

- Domain Analysis
Genomic Data Structures

Strings

..TTGAATTACATG..
|||  |||
GAA--ACA

Alignment
Narzisi et al. (2014) Nature Methods
Lee & Schatz (2012) Bioinformatics

Autism Genetics
Iossifov et al. (2014) Nature
Fang et al. (2014) Genome Medicine

Trees

Suffix Trees
Marcus et al. (2014) Bioinformatics
Trapnell & Schatz (2009) Parallel Computing

Microbial Diversity
Donia et al. (2011) PNAS
Schatz & Phillippy (2012) GigaScience

Graphs

String Graphs
Narzisi et al. (2014) Lecture Notes in CS.
Koren et al. (2012) Nature Biotechnology

Plant Biology
Schatz et al. (2014) Genome Biology
Maron et al. (2013) PNAS
Genomics Graphs

1. Error Correction and Assembly
   *Long Read Single Molecule Sequencing*

2. Pan-Genomics
   *Sequence conservation and divergence*
Genome Complexity

https://en.wikipedia.org/wiki/Genome_size
Sequence Assembly Problem

1. Shear & Sequence DNA

2. Construct assembly graph from overlapping reads

...AGCCTAGGGATGCAGCAGACGT
GGATGCAGCAGACGTCGCATATCCGGTTTGGTCAACCGCGACGAC
CAACCGCGACGACCTCAGCGAA...

3. Simplify assembly graph

---

On Algorithmic Complexity of Biomolecular Sequence Assembly Problem
Assembly Complexity
Often an astronomical number of possible assemblies

Value computed by application of the BEST theorem (Hutchinson, 1975)

\[ W(G, t) = (\det L) \left\{ \prod_{u \in V} (r_u - 1)! \right\} \left\{ \prod_{(u,v) \in E} a_{uv}! \right\}^{-1} \]

L = n x n matrix with \( r_u - a_{uu} \) along the diagonal and -\( a_{uv} \) in entry uv

\( r_u = d^+(u) + 1 \) if \( u = t \), or \( d^+(u) \) otherwise

\( a_{uv} \) = multiplicity of edge from \( u \) to \( v \)

Assembly Complexity of Prokaryotic Genomes using Short Reads.
Assembly Complexity
The advantages of SMRT sequencing
3rd Gen Long Read Sequencing

PacBio RS II

CSHL/PacBio
3rd Gen Long Read Sequencing

PacBio RS II

CSHL/PacBio

Histogram:

0 10k 20k 30k 40k

Images:

- Human
- Plant
- Fish
- Microscopic view
3rd Gen Long Read Sequencing

PacBio RS II

Oxford Nanopore

CSHL/PacBio

CSHL/ONT
Oxford Nanopore MinION

- Thumb drive sized sequencer powered over USB
- Capacity for 512 reads at once
- Senses DNA by measuring changes to ion flow
Nanopore Sequencing

Basecalling currently performed at Amazon with frequent updates to algorithm
Nanopore Readlengths

Oxford Nanopore Sequencing at CSHL
30 runs, 267k reads, 122x total coverage
Between 11 and 73k reads per run!
Mean flow cell: 50 Mbp in 2 days
Max flow cell: 446Mbp in 2 days

- Max: 146,992bp
- 8x over 20kb
- 41x over 10kb
- Mean: 5473bp
- Noise
- Spike-in
Nanopore Alignments

Mean: 6903bp

Alignment Statistics (BLASTN)

Mean read length at ~7kbp
Shearing targeted 10kbp
70k reads align (32%)
40x coverage

Max: 50,900bp
Nanopore Accuracy

Alignment Quality (BLASTN)
Of reads that align, average ~64% identity
Alignment Quality (BLASTN)
Of reads that align, average ~64% identity
“2D base-calling” improves to ~70% identity
Error Correction Methods

Word Analysis of Illumina Reads

Kelly, Schatz, Salzberg (2010)
*Genome Biology*, 11:R116
Error Correction Methods

Quake

Word Analysis of Illumina Reads

Kelly, Schatz, Salzberg (2010)
*Genome Biology, 11:*R116
Error Correction Methods

Quake

Word Analysis of Illumina Reads
Kelly, Schatz, Salzberg (2010)
Genome Biology, 11:R116

PacBioToCA & ECTools

Hybrid Correction Of PacBio using Illumina
Nature Biotechnology. 30:693–700
NanoCorr: Nanopore-Illumina Hybrid Error Correction

https://github.com/jgurtowski/nanocorr

1. BLAST Miseq reads to all raw Oxford Nanopore reads

2. Select non-repetitive alignments
   ○ First pass scans to remove “contained” alignments
   ○ Second pass uses Dynamic Programming (LIS) to select set of high-identity alignments with minimal overlaps

3. Compute consensus of each Oxford Nanopore read
   ○ State machine of most commonly observed base at each position in read

Oxford Nanopore Sequencing and de novo Assembly of a Eukaryotic Genome
Long Read Assembly

S288C Reference sequence
- 12.1Mbp; 16 chromo + mitochondria; N50: 924kbp
Genomic Futures?
iGenomics: Mobile Sequence Analysis
Aspyn Palatnick, Elodie Ghedin, Michael Schatz

The world's first genomics analysis app for iOS devices

BWT + Dynamic Programming + UI

First application:
• Handheld diagnostics and therapeutic recommendations for influenza infections
  • Coming soon to the App Store

Future applications
• Pathogen detection
• Food safety
• Biomarkers
• etc..
Genomics Graphs

1. Error Correction and Assembly
   *Long Read Single Molecule Sequencing*

2. Pan-Genomics
   *Sequence conservation and divergence*
Pan-Genome Alignment & Assembly

Time to start considering problems for which N complete genomes is the input to study the “pan-genome”
- Available today for many microbial species, near future for higher eukaryotes

Pan-genome colored de Bruijn graph
- Encodes all the sequence relationships between the genomes
- How well conserved is a given sequence?
- What are the pan-genome network properties?

SplitMEM: A graphical algorithm for pan-genome analysis with suffix skips
Graphical pan-genome analysis

Colored de Bruijn graph

- Node for each distinct kmer
- Directed edge connects consecutive kmers
- Nodes overlap by k-1 bp

AGAAGTCC
ATAAGTTA

de Bruijn, 1946
Idury and Waterman, 1995
Pevzner, Tang, Waterman, 2001
Graphical pan-genome analysis

Colored de Bruijn graph
• Node for each distinct kmer
• Directed edge connects consecutive kmers
• Nodes overlap by k-1 bp

More specifically:
• We aim to build the compressed de Bruijn graph as quickly as possible without considering every distinct kmer
Suffix Trees

Elegant, widely used full text index

- Rooted, directed tree with a leaf corresponding to each suffix
- Path from root to leaf $i$ spells suffix $S[i \ldots n]$
- Each internal node has at least two distinct children except possibly the root
- Special *suffix links* navigate between internal nodes corresponding to consecutive substrings ($x\alpha \rightarrow \alpha$) without returning to root

Many important search problems can be solved in linear time and space

Linear pattern matching algorithms.

On-line Construction of Suffix Trees
Maximal Exact Matches (MEMs)

Definition:
A MEM is an exact match within a sequence that cannot be extended left or right without introducing a mismatch.

...XGATTACAW...   ...YGATTACAZ...

Key Properties:
• MEMs are internal nodes in the suffix tree that have left-diverse descendants.
• Have descendant leaves that represent suffixes with different characters preceding them

• Linear-time traversal of suffix tree to identify MEMs.
MEMs to compressed de Bruijn Graphs

T G C A C ... G G C A A
Overlapping MEMs

TGCCATC GCCAACC CATG
TGCCATC GCCAACC CATG
SplitMEM Sketch

1. Find nodes representing repeated sequences
   1. Build suffix tree of genome
   2. Mark internal nodes that are MEMs, length $\geq k$
   3. Preprocess suffix tree for LMA queries
   4. Determine repeat-nodes of compressed de Bruijn graph by decomposing MEMs and extracting overlapping components, length $\geq k$

2. Finalize graph with nodes and edges of unique sequences
Split MEMs to de Bruijn Graph

Find deepest MEM in suffix tree.
Split MEMs to de Bruijn Graph

Traverse suffix link.
Look for MEM as ancestor.
Split MEMs to de Bruijn Graph

Traverse suffix link.
Look for MEM as ancestor.
Split MEMs to de Bruijn Graph

Traverse suffix link.
Look for MEM as ancestor.
Found MEM as ancestor. Decompose. Remove embedded MEM (suffix links). Find next embedded MEM.
### Suffix Skips

**Genome:** babab

<table>
<thead>
<tr>
<th>Suffix skips 0</th>
<th>Suffix skips 1</th>
<th>Suffix skips 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>(dist = 1; suffix links)</td>
<td>(dist=2)</td>
<td>(dist=4)</td>
</tr>
</tbody>
</table>

- **Suffix Skips 0:**
  - Skip 0 characters in log(0) steps instead of 0 suffix links.

- **Suffix Skips 1:**
  - Skip 1 characters in log(1) steps instead of 1 suffix links.
  - Pointer jumping technique: \( n \rightarrow ss[i] = n \rightarrow ss[i-1] \rightarrow ss[i-1] \)

- **Suffix Skips 2:**
  - Skip 2 characters in log(2) steps instead of 2 suffix links.

---

"Skip c characters in log(c) steps instead of c suffix links.

- Pointer jumping technique: \( n \rightarrow ss[i] = n \rightarrow ss[i-1] \rightarrow ss[i-1] \)"
Microbial Pan-Genomes

**E. coli (62) and B. anthracis (9) pan-genome analysis**
- Analyzed all available strains in Genbank
- Space is linear in the number of genomes
- Time is $O(n \log g)$ where $g$ is the length of the longest genome
  - Linear time for most practical applications

- Many possible applications:
  - Identifying “core” genes present in all strains
  - Characterizing highly variable regions
  - Cataloging sequences shared by pathogenic varieties

![62 strain E. coli Pan-Genome Node Sharing](image)
The Rise of Pan-Genomics

Human Pan-Genomics
• We now have the capacity to consider the pan-genome structure of the human population and other high value species

• Already the current human reference genome has “alternate” sequence paths representing major differences between the different ethnicities (haplotype groups)

• However, virtually none of existing genomics algorithms operate on reference graphs, creating a major opportunity for research:
  • New and interesting CS problems
    • Online graph construction, searching, annotating, visualizing…
  • New and interesting biology
    • Detailed analysis of mutation, disease, and evolution

Extending reference assembly models
Interfacing CS & Biology

Theory & Programming Languages
• How can we efficiently search & analyze genomic data?
• How do natural systems use abstraction or recursive processing?

Systems
• How do we scale to exascale or zettascale genomic data?

Information Security
• How do we balance the benefits of sharing genomic data with potential privacy abuses?

Machine Learning & Data Intensive Computing
• How do we learn from high dimensional biological data?

Language & Speech Processing
• How do we recognize important features of sequences and other bio-molecular data?

Robotics, Vision & Graphics
• How do we integrate and model molecular with behavioral data?
Understanding Genome Structure & Function

**Genomics is a rich field for computer science research**

- Opportunities across the entire data science spectrum from sensors & data systems, through algorithmics and machine learning

**Sequencing Algorithmics**

- Long reads and other sequencing technologies are giving us great power to look into genomes across the tree of life
- With these advances, expect the rise of graph-based pan-genomics giving us new insights into the origins of disease, the processes of development, and the forces of evolution

*Also very interested in teaching the next generation of undergraduate and graduate students*
Acknowledgements

**Schatz Lab**
- Rahul Amin
- Eric Biggers
- Han Fang
- Tyler Gavin
- James Gurtowski
- Ke Jiang
- Hayan Lee
- Zak Lemmon
- Shoshana Marcus
- Giuseppe Narzisi
- Maria Nattestad
- Aspyn Palatnick
- Srividya
- Ramakrishnan
- Rachel Sherman
- Greg Vurture
- Alejandro Wences

**CSHL**
- Hannon Lab
- Gingeras Lab
- Jackson Lab
- Hicks Lab
- Iossifov Lab
- Levy Lab
- Lippman Lab
- Lyon Lab
- Martienssen Lab
- McCombie Lab
- Tuveson Lab
- Ware Lab
- Wigler Lab
- IT & Meetings Depts.
- Pacific Biosciences
- Oxford Nanopore
Thank you

http://schatzlab.cshl.edu
@mike_schatz