Compressive Acceleration in Biological Data Science

Noah M. Daniels
Sequencing: Ancient History

Sanger, Gilbert

Smith-Waterman

Human Genome Project

1000 Genomes Project

1970

1980

1990

2000

2010

Needleman-Wunsch

BLAST (Myers, et al.)

Human Genome Project

1000 Genomes Project
“It costs more to analyze a genome than to sequence a genome.” – David Haussler

“Researchers are increasingly turning to cloud computing…”
A bigger cloud?

Still constrained by Moore
Genomic redundancy

D. simulans
D. sechellia
D. melanogaster
D. yakuba
D. erecta

Unique data
Why is this hard?

Standard compression doesn’t respect the data structures!
Search needs inexact matches

Compute in compressed space
How does our compression work?

Lossless compression!
How does our search work?

1. Coarse Search

Query Sequence

Coarse Database

Candidate Hits

2. Sequence reconstruction

Candidates

3. Fine Search

Results
Why should this be faster?

Sequences in the coarse database are:

• long enough to not be degenerate
• dissimilar to each other
• similar to the sequences they represent

BLAST-searchable, but small
Is compressive search faster?
Simulated data growth

- **Search time (seconds)**
- **Size of simulated yeast genome**
- **20% mutation rate**
- **Compressive-accelerated BLASTP**
- **BLASTP**
Small molecule search
Small molecule search

[Chemical structures of caffeine and related molecules]
Metagenomic search