Seminar on Algorithms for Deep Sequencing Data

سميיר באלגוריתמים לנתוני ריצוףعمוק

0368-4198-01

Prof. Ron Shamir

School of Computer Science

Fall 2018
Outline

• A little bit of biology
• Gene expression
• Sequencing
• Deep sequencing – NGS
• About the seminar
• Your opportunity to ask lots of questions!!!
A little bit of biology
Gregor Mendel laws of inheritance, “gene” 1866

Watson and Crick DNA Discovery 1953

Genome Project 2003
DNA and Chromosomes

- **DNA**: 4 bases molecule: ACGT

- **Chromosome**: contiguous stretch of DNA

- **Genome**: totality of DNA material
Genes: Recipes for Proteins

- **Gene**: a DNA segment that specifies the sequence of a protein.
- **RNA**: copy of DNA of a gene; “manufacturer instructions” for a protein
DNA $\rightarrow$ RNA $\rightarrow$ protein

The hard disk $\uparrow$

One program $\uparrow$

Its output $\uparrow$

transcription $\uparrow$

translation $\uparrow$

© Ron Shamir
Biology and Computation
One of many computational challenges in the Human Genome project:

Assemble a puzzle of 27 million pieces
Complexity

- ~3,000,000,000 letters in the genome
- 2,278,100 letters in the Bible
- => one genome = a stack of ~ 1,000 Bibles

- ~20,000 genes in the genome
- Hard to identify
- Harder to figure their function
- Even harder to figure how they work together
Enter Bioinformatics

- The marriage of CS and Biology
- Responds to the explosion of biological data, and builds on the IT revolution
Biology is becoming an information science.
Expression profiling
The cell as a busy chef

DNA → RNA → protein

The expression profile of the cell: which genes produce mRNA and at what quantities

• Profiles vary dramatically between cells, tissues, situations, developmental stages...
• To understand life and disease we must understand gene regulation

© Ron Shamir
DNA chips / Microarrays

- Simultaneous measurement of expression levels of all genes.
- Perform $10^5$-$10^6$ measurements in one experiment
- Allow global view of cellular processes.
The Raw Data

Entries of the Raw Data matrix: Ratios/absolute values/…

- **expression pattern** for each gene
- Profile for each experiment /condition/sample/chip

Needs normalization!
GEO

Gene Expression Omnibus

GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles.

<table>
<thead>
<tr>
<th>Getting Started</th>
<th>Tools</th>
<th>Browse Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overview</td>
<td>Search for Studies at GEO DataSets</td>
<td>Repository Browser</td>
</tr>
<tr>
<td>FAQ</td>
<td>Search for Gene Expression at GEO Profiles</td>
<td>DataSets: 4348</td>
</tr>
<tr>
<td>About GEO DataSets</td>
<td>Search GEO Documentation</td>
<td>Series: 103543</td>
</tr>
<tr>
<td>About GEO Profiles</td>
<td>Analyze a Study with GEO2R</td>
<td>Platforms: 18984</td>
</tr>
<tr>
<td>About GEO2R Analysis</td>
<td>Studies with Genome Data Viewer Tracks</td>
<td>Samples: 2682501</td>
</tr>
</tbody>
</table>

~2.7 million expression profiles
All publicly available, well organized
A vast, underutilized resource.
Sequencing
How do we read DNA?
How do we read DNA?

• We replicate it
How do we read DNA?

• We shred it

• We pick many short segments (of unknown location) for reading
Reading DNA until 2008

Sanger Sequencing

Sanger sequencing uses ddNTPs (dideoxynucleotide triphosphates) which do not have a free 3' OH mixed in with dNTPs. Whenever the DNA polymerase incorporates a ddNTP it won’t be able to add any other nucleotides. Then gel electrophoresis is used to separate the DNA

https://www.youtube.com/watch?v=593zWZNwbJI
Reading short DNA

- Use replication machinery with colored bases
- Take pictures of massively parallel reaction
- 10 million reads of 30 bp per day & $1000
Reading short DNA

- Use replication machinery with colored bases
- Take pictures of massively parallel reaction
- 100 million reads of 100 bp per day & $100s
600B bases per day!!
Whole Genome

Whole genome sequencing is a random approach that harnesses the power of NGS to completely sequence an entire genome.

Applications

- *De novo* assembly and analysis of novel genomes
- Comparative analysis with other known genomes
- Update gene annotation

<table>
<thead>
<tr>
<th>Sequencing Product Description</th>
<th>Read Chemistry</th>
<th>Price per Library Prep</th>
<th>Price per Gbase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole genome sequencing</td>
<td>PE 2×150</td>
<td>$150</td>
<td>$40</td>
</tr>
</tbody>
</table>

**SPECIAL!** Human genome sequencing (PE 2 x 150) at 30x coverage (includes library preparation):

- $1,300 per sample (> 100 samples)
- $1,500 per sample (> 10 samples)
- $2,000 per sample (< 10 samples)
Get the most comprehensive genetic testing service there is.

Make more informed decisions about your health, learn about your ancestry, and much more.

Ready to order?

LET’S GO!

Want to learn more?
Whole GenomeM: Whole Genome Sequencing with mtDNA - Full DNA Analysis for Genetic Diseases

$599.00  $899.00  You save $300.00

ADD TO CART
Example (Itzik Pe’er)

The genome is:

```
TTATGGTCGGTGAGTGTGACTGGTGTTGTCTAA
```

The reads are:

```
GGTCGGTGAG
TGAGTGTGAC
TGGTGGTGTC
TGACTGGTTT
AATGGTCGGT
GAGTGTGACT
```
```
Example

The genome is:

```
TTATGGTCGGGTAGTGTGACTGGTGTTGTCTAA
```

```
GGTCGGTGAG
```

The reads are:

```
GGTCGGGTAG
TGAGTGTGAC
TGGTGTTGTC
TGACTGGTTT
AATGGTCGGT
GAGTGTGACT
AAAAAAA
```
Example

The genome is:

```
TTATGGTGTCGGTGAGTGTGACTGGTGTTGGTCTAA
```

The reads are:

```
GGTCGGTGAG
TGAGTGTGAC
TGGTGTTGTC
TGACTGGTTT
AATGGTGTCGG
GAGTGTGACT
AAAAAAAAAA
```
Example

The genome is:
TTATGGTCCGTTGAGTGTGAC\textcolor{red}{TGGTGTGGTCTAA}\\ |
|
|
|
|
|
|
|
|
TGTTGTTGTC

The reads are:
GGTCGGTGGAG
TGAGTGTGAC
\textcolor{red}{TGGTGTGGTGC}
TGACTGGTTTT
AATGGTCCGTT
GAGTGTGACT
AAAAAAAAAAAAA
Example

The genome is:

TTATGGTCGGTGAGTGTGACTGGTGTTTGTCTAA

TGACTGGTTT

The reads are:

GGTCGTTGAG
TGAGTGTGAC
TGGTGTGGTC
TGACTGGTTT
AATGGTCGGT
GAGTGTGACT
AAAAAAACT
AAAAAAAAAAAAA
Example

The genome is:
TTATGGTCG GTGAGTGTGAC TGGTGTTG TC TAA

The reads are:
GGTCGGTGAG
TGAGTGTGAC
TGGTGTTGTC
TGACTGGTTT
AATGGTCG GTG
GAGTGTGAC T
A A A A A A A A A A

Mapping problem:
Given reference and query, find best match(es)

Mapping problem (2):
Given reference and LOTs of queries, find best match(es) for each
Example

The genome is:

```
TTATGGTCGGTGAGTGTGACTGGTGTTGTCTAA
GGTCGGTGAG
TGAGTGTGAC
TGGTGTTGTC
TGACTGGTTT
AATGGTCGGT
GAGTGTGACT
```

AAAAAAAAAAAAA
Sequence assembly
What if we have no reference?
Sequence Assembly

- **Input:**
  
  - GGTCGGTGAG
  - TGAGTGACG
  - TGATTGTC
  - TGACTGGTTT
  - AATGGTCGGT
  - AAAAAAAAA

- **Output:**
  
  - ATATGGTCGGTGAGTGTGACTGGTTGTCTAA

Assembly problem: Given many reads from an unknown sequence, reconstruct it.
The de Bruijn graph of order $k$:

- **Vertex** = $k$-mer $(x_1,...,x_k)$  \[ |V| = |\Sigma|^k \]

- **Directed edge for overlap of** $(k-1)$:
  \[(x_1,x_2,...,x_k) \rightarrow (x_2,...,x_k,x_{k+1})\]  \[ |E| = |\Sigma|^{k+1} \]

- $\forall v \ d^+(v) = d^-(v) = |\Sigma|$  

We use the name "de Bruijn graph" also for subgraphs induced by a subset of the vertices.
de Bruijn graph for $k=4$, 0/1 alphabet

de Bruijn sequence - Euler cycle on $G$ - includes each k-mer once 0000110010111101
de Bruin Graph

Nodes: k-tuples
Idea: de Bruin Graph

Edges: \((k+1)-\text{tuples}\)
Sequence $\iff$ Path in Graph
Assembly Using de-Bruijn Graphs

Input:
GGTCGGTGAG
TGAGTGTGAC
TGGTGTGTGTC

1. Turn reads to paths
GGTC→GTCG→TCGG→CGGT→GGTG→GTGA→TGAG
TGAG→GAGT→AGTG→GTGT→TGTG→GTGA→TGAC
TGGT→GGTG→GTGT→TGTT→GTTG→TTGT→TGTC
Assembly Using de-Bruijn Graphs

1. Turn reads to paths
2. Merge paths
Assembly Using de-Bruijn Graphs

1. Turn reads to paths
2. Merge paths
3. Resolve error “bubbles”
4. Resolve cycles (repeats)
Overview of Selected HTS Applications

Publication date of a representative article describing a method versus the number of citations that the article received. Methods are colored by category, and the size of the data point is proportional to publication rate (citations/months). The inset indicates the color key as well the proportion of methods in each group. For clarity, seq has been omitted from the labels.

Reuter et al. High-Throughput Sequencing Technologies Mol Cell 15
Gene Structure

Introns vs Exons

https://www.youtube.com/watch?v=_asGjfCTLNE
Some terminology you are likely to encounter

Genome, transcriptome, exome

Assembly, contig

Reverse complementarity

K-mer counting, canonical k-mers

Sequencing: “30x” coverage, errors
The seminar
Guidelines

• You will sometimes need to **dig deeply** for the methods: supplements (on journal websites), previous papers,..

• See seminar website for resources

• (re)start with the **basics**: definitions, examples

• Papers contain more than you can cover: **Select** your presentation focus wisely
Guidelines (2)

• Provide **intuition** and examples to motivate your method / theorems
• Add something **original** that you thought of (and don’t hide that!)
• Focus more on the **algorithms** than on the **results** (rule of thumb: 80-20)
Planning your presentation

• Start: 5:10, Break 6-6:10, Talk End: 6:40, followed by 5 min for questions, then open discussion
• Use mostly slides, and the board sparingly
• Rehearse your talk!
• Make contingencies in case you’re out of time
• In the end, summarize the paper, repeating the main results. Discuss strengths, weaknesses, steps ahead.
Planning your presentation (2)

• When two students present the same paper together:
  – Make sure each of you understand everything!
  – Split the presentation evenly in terms of time and contents
  – Switching multiple times between you is fine (and often a good idea)
  – Coordinate so that the presentation is seamless.
The questionnaire

- Prepare a short (4-5 item) questionnaire on the paper
- Level should be basic, but require reading the paper
- Print in advance and distribute it to students at the end of the seminar
- Students will bring in their answers next week, and you will grade them and return to me by the following Tuesday.
קביעת הציון הסופי:

• הבנת החומר: 35%
• הצגת החומר: 35%
• בחירה טובה איזה חומר להציג: 10%
• השתתפות פעילה בסמינר (shivohot ודקיפシェאלות): 20%
• בonus על מקורות: 10%
• חירגה מהدافع: 10% -!!
• kııımת חובת זוכחות בסמין