Similarity Search with DNA

Using DNA to do things silicon computers have traditionally done
Overview

What is similarity search?
Why would you want to use DNA?
How do you use DNA?
How could we make this better?
How do you use Cas9 to perform similarity search?
Musings on the intersection of Comp Bio, Molecular Computing, SynBio
What is similarity search?

Input

![Input Image]

Output

![Output Images]
Why would you want to use DNA?

- Parallelism
- DNA information density
- DNA longevity
- Ease of distributing DNA databases
- Sometimes faster
- Sometimes more energy efficient

In that faint pink smear is ~10TB of data
Similarity Search

Curse of high dimensionality:
Exact indexing schemes in high-dimensionality spaces are no better than a costly linear or “brute force” search, which is infeasible for large databases.

Tradeoff:
Rather than finding exact nearest neighbors, our goal is to maximize the number of near neighbors retrieved while minimizing the number of irrelevant results.
How do we do similarity search with DNA?

A. Document similarity as geometric space
How do we do similarity search with DNA?

A. Document similarity as geometric space

B. Similarity-preserving DNA encoding

- Neighboring documents
- Stable DNA hybridization

- Query
- Target

- Distant documents
- Unstable DNA hybridization
How do we do similarity search with DNA?

A. Document similarity as geometric space

B. Similarity-preserving DNA encoding

C. Database

D. Synthesized index

E. Synthesized query

F. Query hybridized with index

G. Index filtered by query

H. Database filtered by query

Encoding/Synthesis

t-SNE 1

t-SNE 2

ID Features

Neighboring documents → Stable DNA hybridization

Distant documents → Unstable DNA hybridization

Query

Target

Query Target

Stable DNA hybridization

Unstable DNA hybridization
How do we do similarity search with DNA?

A. Document similarity as geometric space

B. Similarity-preserving DNA encoding

C. Database
D. Synthesized index
E. Synthesized query
F. Query hybridized with index
### A. Document similarity as geometric space

![Image of document similarity as geometric space](image1.png)

### B. Similarity-preserving DNA encoding

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Query hybridized with index</td>
</tr>
<tr>
<td>2</td>
<td>Stabilize DNA hybridization</td>
</tr>
<tr>
<td>3</td>
<td>Unstable DNA hybridization</td>
</tr>
</tbody>
</table>

#### Diagram Elements

- **t-SNE 1**: Neighboring documents
- **t-SNE 2**: Distant documents
- **Query**: Original DNA
- **Target**: Modified DNA

### C. Database

- **ID**: Identification
- **Features**: Characteristics

### D. Synthesized index

- **ID**: Identification
- **Features**: Characteristics

### E. Synthesized query

- **ID**: Identification
- **Features**: Characteristics

### F. Query hybridized with index

- **Biotin**: Marker

### G. Index filtered by query

### H. Database filtered by query

- **Remove unbound strands**
- **Hybridization**
- **User digest**

#### Notes

- **Magnetic beads**
- **PCR (10-15 cycles)**
- **Linear PCR (3 cycles)**
- **Linear PCR (1 cycle)**

### Encoding/Synthesis

- **Annealing**
- **Bead filtration**
- **Sequencing/decoding**

### Sequencing/decoding

- **Illumina ligation, followed by sequencing**
It works!

A. Distribution of similarity across read depths

B. Retrieval as a function of read depth

C. Sets of retrieved images for select read depth thresholds

<table>
<thead>
<tr>
<th>Number retrieved</th>
<th>100-NN recall</th>
<th>Top 5 results (sorted by similarity)</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>0.01</td>
<td><img src="image1.png" alt="Image 1" /></td>
</tr>
<tr>
<td>58</td>
<td>0.05</td>
<td><img src="image2.png" alt="Image 2" /></td>
</tr>
<tr>
<td>141</td>
<td>0.10</td>
<td><img src="image3.png" alt="Image 3" /></td>
</tr>
<tr>
<td>1,831</td>
<td>0.50</td>
<td><img src="image4.png" alt="Image 4" /></td>
</tr>
<tr>
<td>315,736</td>
<td>0.94</td>
<td><img src="image5.png" alt="Image 5" /></td>
</tr>
</tbody>
</table>

The proportion of the entire dataset that must be retrieved (y-axis) to retrieve a certain proportion of the 100 most similar images (x-axis).
How can we improve?

• More energy efficient
• Faster

...what if we used Cas9?
Using Cas9

Hybridization and Bead Extraction

**Cas9**

Using Cas9

We sequence every strand that’s cut by Cas9
Using Cas9

We sequence every strand that's cut by Cas9

Hard to do in wet lab

We sequence every strand that's cut at BOTH Cas9 regions...AND logic!
Using Cas9

We sequence every strand that's cut by Cas9.

Hard to do in wet lab

We sequence every strand that's cut at BOTH Cas9 regions...AND logic!

What if we allow for nested queries??
The lines are still being drawn

My two cents for people in any of these three fields:

- If you’re a computation-centered person, get comfortable talking to wet lab-centered people
- vice-versa
- Bigger computational AND molecular toolboxes tend to make it easier to design experiments
- We need more tools
The lines are still being drawn on the map of how these fields intersect.

My two cents for people in any of these three fields:

- If you’re a computation-centered person, get comfortable talking to wet lab-centered people
- vice-versa
- Bigger computational AND molecular toolboxes tend to make it easier to design experiments
- We need more tools

Comp Bio ∩ Molecular Computing ∩ SynBio

Cries in Synthetic Biologist

Kyle 🪐 🦈 🐸 @KyleMorgenstein · 23h
how the hell did we make planes before CAD??

14:24 · 5/20/21 · Twitter Web App
5 Retweets 31 Likes

Sebastian S. Cocioba @ATinyGreen · 5h
Repeating to @p_maverick_b
What’s the slide-rule of synthetic biology tho?

Patrick Boyle @p_maverick_b · 5h
Back in my day we tagged everything with GFP whether you needed it to glow or not
Things to think about

• What are some tools you’d like to see developed?
• What are things you’d like to see standardized?
• Are there times when having a deeper background in a different field (i.e., biology) would have helped you?
• Anything other wishes for the future?