Similarity Search with DNA Using DNA to do things silicon computers have traditionally done

Lee Organick

5-24-21

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

Output





Perú: La Sierra hasta la Selva . musingsfromthesaddle.com



Guided Backcountry Foraging | the ... hehungryforagencom - In stock



pay later | Quadpay quadpay.com



Similar Images



Hammastunturi Wilderness | 5 days of ... flick.com





246 Takemoto Photos and Premium High ... gettyimages.dk



South Fork Camp 2 | Bob Bukantis | Flickr flickscom



Backpacking Copper Ridge Loop in North ... sestfielded.com



Six Days in Olympic Nati... dauchebacker.com

cynhw.com



Upper Crystal Rainier - ... North Cascades Trip - Back to the PCT ...

reneeandtimhike.com

The North Face Stormbreak 2 Reviews ... railspace.com



Twin Lakes - Leadville Primitive | The Dyrt





kaodronides.com



Adventure Photography Portfolio brand or cookphot agruphy, cam-





Backpacking Alone Is for Everyone ... awiascycles.com





567 Tarp Tent Photos and Premium High ...

orthringes.com



Horn Fork Basin Dispersed Camping | The... thedyrt.com





Outdoor gear . pinterest.com





thedyrt.com



Stone Glacier - SkyAir ULT stanegladier.com



isle Royale National Park - Kao Chronicles



Hike In Camping Experience. hipcamp.com

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



*Content-Based Similarity Search in Large-Scale DNA Data Storage Systems. Bee et al. BioRXiv 2020

Similarity Search

A. Document similarity as geometric space



Curse of high dimensionality:

Exact indexing schemes in highdimensionality spaces are no better than ostly linear or "hrute force" search



Fradeoff:

ather than the ding exact nearest goal is to maximize the neighbors, our number of near neighbors retrieved while minimizing the number of irrelevant results





A. Document similarity as geometric space



A. Document similarity as geometric space



Neighboring documents





Query





Distant documents

A. Document similarity as geometric space



Neighboring documents





Query





Distant documents













It works!

A. Distribution of similarity across read depths

B. Retrieval as a function of read depth





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)



C. Sets of retrieved images for select read depth thresholds





How can we improve?

- More energy efficient
- Faster

...what if we used Cas9?

Hybridization and Bead Extraction









We sequence every strand that's cut by Cas9

















Comp Bio ∩ Molecular Computing ∩ SynBio

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

Computational Biology

Synthetic Biology

Molecular Computing



Comp Bio ∩

The lines are still being draw

My two cents for people in a these three fields:

- If you're a computation-c person, get comfortable t wet lab-centered people
- vice-versa
- **Bigger computational AN** molecular toolboxes tend make it easier to design experiments
- We need more tools



Patrick Boyle @p_maverick_b

Cries in Synthetic Biologist



how the hell did we make planes before CAD??

 $\uparrow \downarrow$

14:24 · 5/20/21 · Twitter Web App

5 Retweets 31 Likes

 \bigcirc 3

 \mathcal{Q}



Sebastian S. Cocioba @ATinyGre... · 5h ···· Replying to @p_maverick_b

 \bigcirc

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

Tweet

...

ட

 \mathbb{C} <u>,</u>↑,

ing n SynBio

Synthetic Biology

ar Computing



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?