Sequence-based Searching For SynBioHub Using VSEARCH

Eric Yu
Chris Myers
University of Utah
Salt Lake City, UT, USA
eric.j.yu@outlook.com,myers@ece.utah.edu

1 INTRODUCTION

With the massive growth of community designed parts, open-source repositories such as SynBioHub [2] have become increasingly popular among synthetic biologists as a convenient way to store and share their genetic designs online. However, the sheer size of such repositories make it difficult to simply browse for the desired parts. The reference instance (https://synbiohub.org) contains over 100,000 publicly available parts in its repository, not counting the various private repositories added by users. Currently, users can only find a part based on a keyword in the part's description, or through various filters such as date of creation, creator, or collection. However, prior to this work, it was not possible in SynBioHub to search for similar sequences.

Well-known tools such as *BLAST* already exist, but are not well-suited for sequence-based searching, as its use of a local alignment algorithm (which aligns a substring of the query sequence to a substring of the target sequence) is more suited for finding patterns between divergent sequences within other domains, which can lead to false positives. *VSEARCH* [3], an open-source alternative to the *USEARCH* [1] tool, uses a more suitable global alignment algorithm, which is more effective at comparing similarities over entire sequences.

VSEARCH was implemented in SynBioHub through *SBOLExplorer* [4], a tool that enhances search by applying PageRank, clustering analysis, and other techniques. Users can either use sequence-based searching through SynBioHub's web interface (Figure 1), or through SynBioHub's API by sending GET requests using *curl* or other languages to get a JSON formatted output instead of a page of results. Various options allow the user to tweak their their search results, which may affect the runtime of *VSEARCH*.

2 RESULTS

With the addition of a new sequence search page at (https://synbiohub.org/sbsearch), users can search by either entering their sequence into a text box or uploading a FASTA or FASTQ file. Table 1 shows the various options that users can adjust when sequence searching.

For users who prefer the command line, SynBioHub's API documentation (https://synbiohub.github.io/api-docs/#search-endpoints) provides instructions to write a GET request using either curl, Python, or JavaScript. An example of a Python script querying for exact matches of a sequence via file upload is shown below:

```
import requests
response = requests.get(
    'http://localhost:7777/search/file_search='
    +'%2FUsers%2Fericyu%2FDownloads%2Fseq.fsa&'
    +'search_exact=true&',
    params={'X-authorization': '<token>'},
    headers={'Accept': 'text/plain'},
)
print(response.status_code)
print(response.content)
```

After submitting the GET request to SynBioHub, users will receive a JSON-formatted output similar to the result below:

```
[{"type":"http://sbols.org/v2#ComponentDefiniti
on","uri":"https://synbiohub.org/public/igem/
BBa_J06480/1","name":"BBa_J06480",
"description":"R0079.B0015",
"displayId":"BBa_J06480","version":"1"}]
```

3 DISCUSSION

Further work is being continually done to add support for more options when using sequence searching through SynBioHub. Additionally, sequence searching will be used as part of the Synthetic Biology Knowledge System (SBKS) that leverages existing data repositories and publications to create a single interface in order to deliver effective and efficient access to collectively available information.

4 METHODS

VSEARCH was made accessible in *SBOLExplorer* through an endpoint implemented in Python using the Flask package, allowing SynBioHub's NodeJS backend to query *SBOLExplorer*.

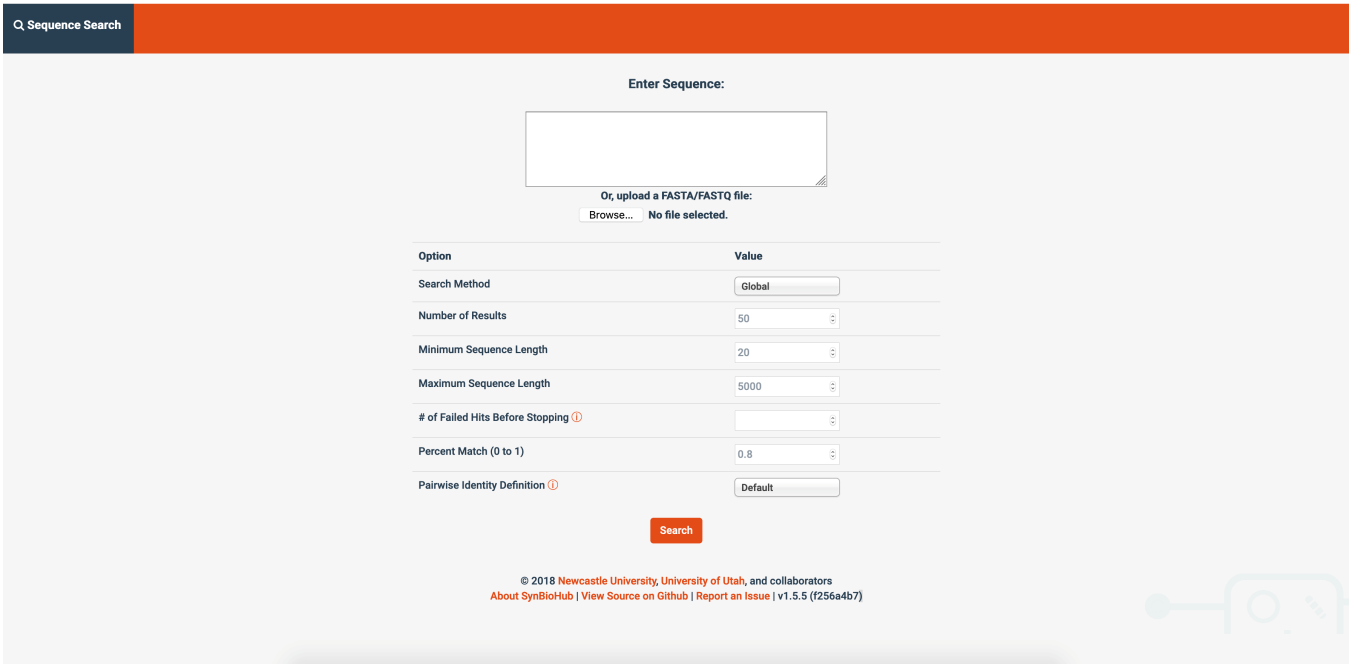

Figure 1: Sequence search tool on SynBioHub

Table 1: Description of search options available to users.

Option Name	Description	Default Value
Search Method	Ability to search by exact match or by some identity threshold (see "Pairwise Identity Definition")	Global
Number of Results	Number of hits before stopping search. Note that a higher number of results will lead to an increase in search time.	50 results
Minimum/Maximum Sequence Length	All sequences below or above the base pair threshold specified will be excluded from the database for comparison.	Min: 20bp Max: 5000bp
# of Failed Hits Before Stopping	Number of false matches before stopping search.	N/A
Percent Match	Float between 0 and 1 specifying percentage identity to query sequence. Anything below the threshold will not be included in the search results.	0.8
Pairwise Identity Definition	Formula to calculate percentage match between query and target sequence.	Edit distance excluding terminal gaps

REFERENCES

[1] Edgar, R. Search and clustering orders of magnitude faster than blast. *Bioinformatics (Oxford, England) 26* (10 2010), 2460–1.

[2] McLaughlin, J. A., Myers, C. J., Zundel, Z., Misirli, G., Zhang, M., Ofiteru, I. D., Goñi-Moreno, A., and Wipat, A. Synbiohub: A standards-enabled design repository for synthetic biology. *ACS Synthetic Biology 7*, 2 (2018), 682–688. PMID: 29316788.

[3] Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* (2016).

[4] Zhang, Michael and Zundel, Zach and Myers, Chris J. SBOLExplorer: Data infrastructure and data mining for genetic design repositories. *ACS Synthetic Biology 8*, 10 (2019), 2287–2294. PMID: 31532640.