GeneQuest

Ashkan Nikfarjam
Tarif Khan

Abstract

Sequence alignment, BLAST searches, and phylogenetic tree building are critical for understanding how genetic sequences work and interact. Sequence alignments aid in identifying conserved regions that can play important roles in biological processes. BLAST and MegaBLAST compare DNA sequences to identify gene similarities. Phylogenetic trees offer a visual picture of evolutionary relationships. (Suzuki et al. 1986). GeneQuest is a web application meant to simplify bioinformatic workflows that include the procedures mentioned above.

Background

In the bioinformatics field working with genetic data can be a lot of work. Performing operations such as querying databases, performing sequence alignments, and building phylogenetic requires a user to use and navigate through many separate tools. Currently researchers rely on tools such as ClustalOmega and NCBI's BLAST and GenBank. These platforms are widely used but their interfaces are often made for experienced users. which beginners may find intimidating. Additionally, each of these tools functions independently. This might make the workflow feel fragmented, complex and it could be time consuming for researchers, particularly for users who are new to bioinformatics and have limited computer abilities. A web application that unifies and simplifies these tools is especially valuable. This is where our project, GeneQuest, comes in. GeneQuest addresses these problems by giving a unified, user-friendly, and simple platform that connects different tools. Our app integrates:

- GenBank search: Allows users to search for and retrieve specific DNA sequences from NCBI's GenBank database.
- BLAST and megaBLAST: Enables users to perform sequence alignments to find matches or similar regions in DNA sequences.
- Phylogenetic tree builder: Offers an easy-to-use platform to visualize evolutionary relationship between species.

- Visual sequence alignment: Allows users to clearly see similarities between sequences and lets users see indels, addition or deletion in genetic sequences. This is especially useful to detect any mutation causing frameshift in a gene. (Chowdhury and Garai, 2017)
- Calculating Conservative regions of aligned sequences: It refers to part of a DNA or protein sequence that remains very similar across different species or even species that belong to the same family(PubMed). It is also particularly important because it helps to identify converging regions for further study. Gen bank offers a utility for calculating and displaying these conservative regions.

By combining these tools into one app, GeneQuest streamlines common bioinformatics workflow

Methods

The first step was analyzing the most demanding tools researchers use for this type of research. It was necessary to find APIs and libraries that could be used for genome search and sequence alinements. NCBI offers 4 different APIs (NCBI):

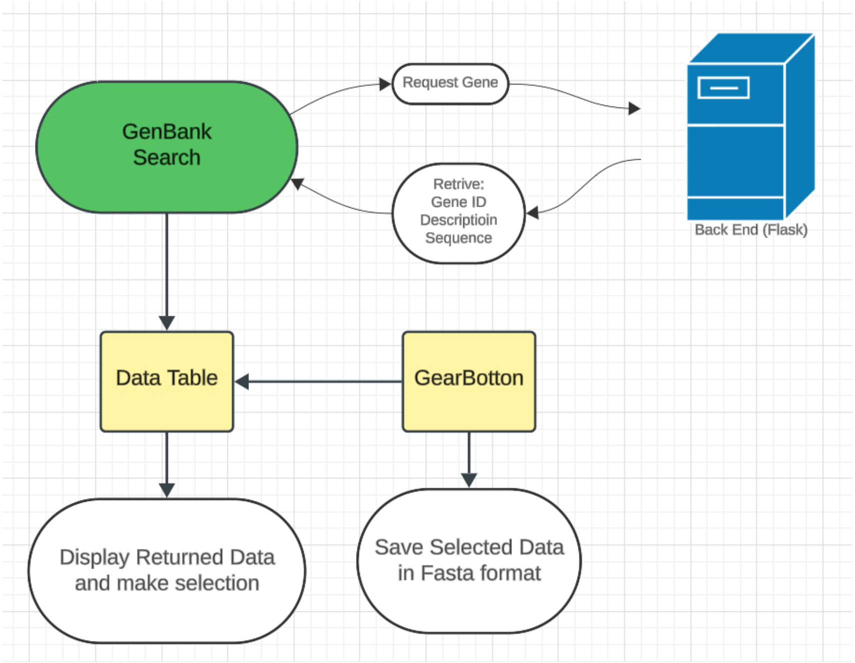
- Entrez Programming Utilities: provide an api for genome search. Our GenBank search utilizes this endpoint for searching genes.
- Blast API: mostly compatible with C++.
- PubMed and PubChem Gateways: for medical and chemistry research.

GeneQuest gathers all the tools needed for genetic research by taking a two layers implementation approach. This design provides modularity, scalability, security, flexibility of components that could be used. Front end utilizes ReactJS, a java script library for building a user interface. A flask server is used in the backend to provide our front end routes and endpoints it needs to request and fetch data from various APIs and libraries such as BioPython and so on.

Front End

React allows developers to break down the UI into reusable, independent components. These components consist of JS functions that return HTML elements.

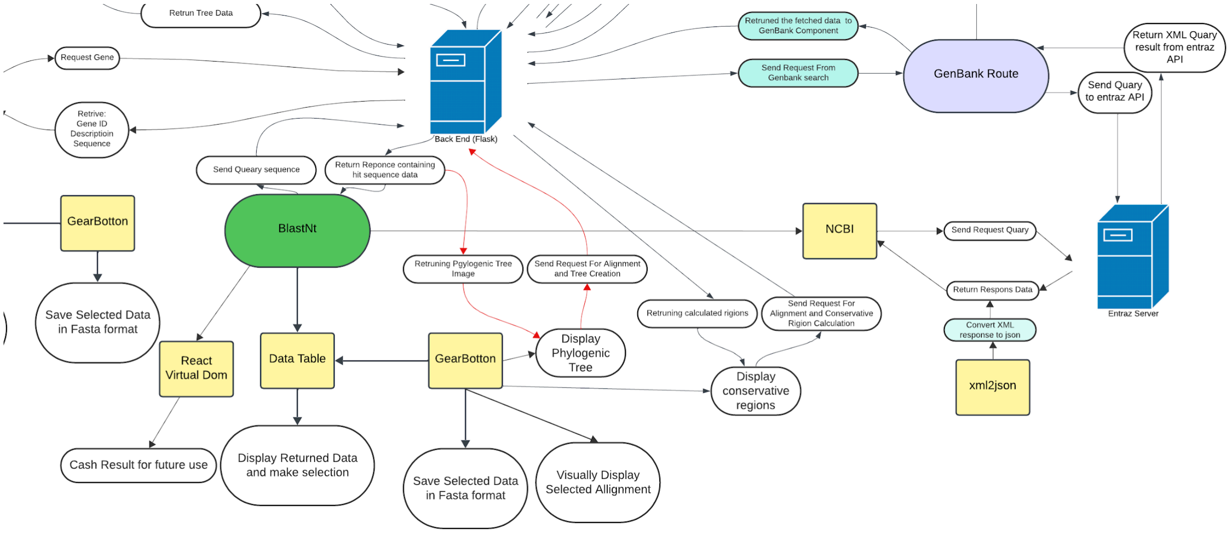
a. GenBank search components: Allows users to query NCBI's GenBank to retrieve specific gene sequences and their variations. Additionally it enables the user to download the fasta file containing extracted genomes.


GenBank Search
Request Gene
Retrive: Gene ID Descriptioin Sequence
Back End (Flask)
Data Table
GearBotton
Display Returned Data and make selection
Save Selected Data in Fasta format

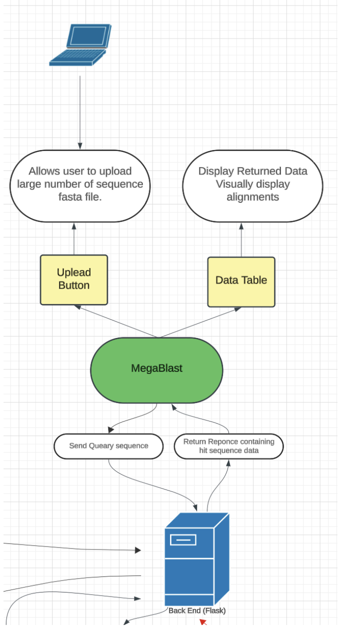
b. BLAST and MegaBLAST: Provides tools for sequence alignment to identify similarities between DNA sequences. In blast, the user is similarly able to save selected hits as a fasta file, visually see the selected alignments, generate a comprehensive phylogenetic tree and display conservative regions of selected sequences for further analysis.

i. Blast uses NCBI.js component as a connector to NCBI blast api "https://blast.ncbi.nlm.nih.gov/Blast.cgi". The app sends a query request in xml format to fetch IDs, descriptions, e-values, and hit sequence data from the NCBI database. Since NCBI is maintained by public funding, it tends to be slower compared to APIs provided by private companies, especially for long sequences. To mitigate the long wait times, the React app caches search results in the browser's localStorage using unique keys. This caching mechanism interacts with the browser's DOM (Document Object Model) environment to store and retrieve data efficiently, reducing the need for repeated requests.

To calculate and display the conservative regions, this component sends a request with information of selected sequences to the back end and only displays the returned result as an image.

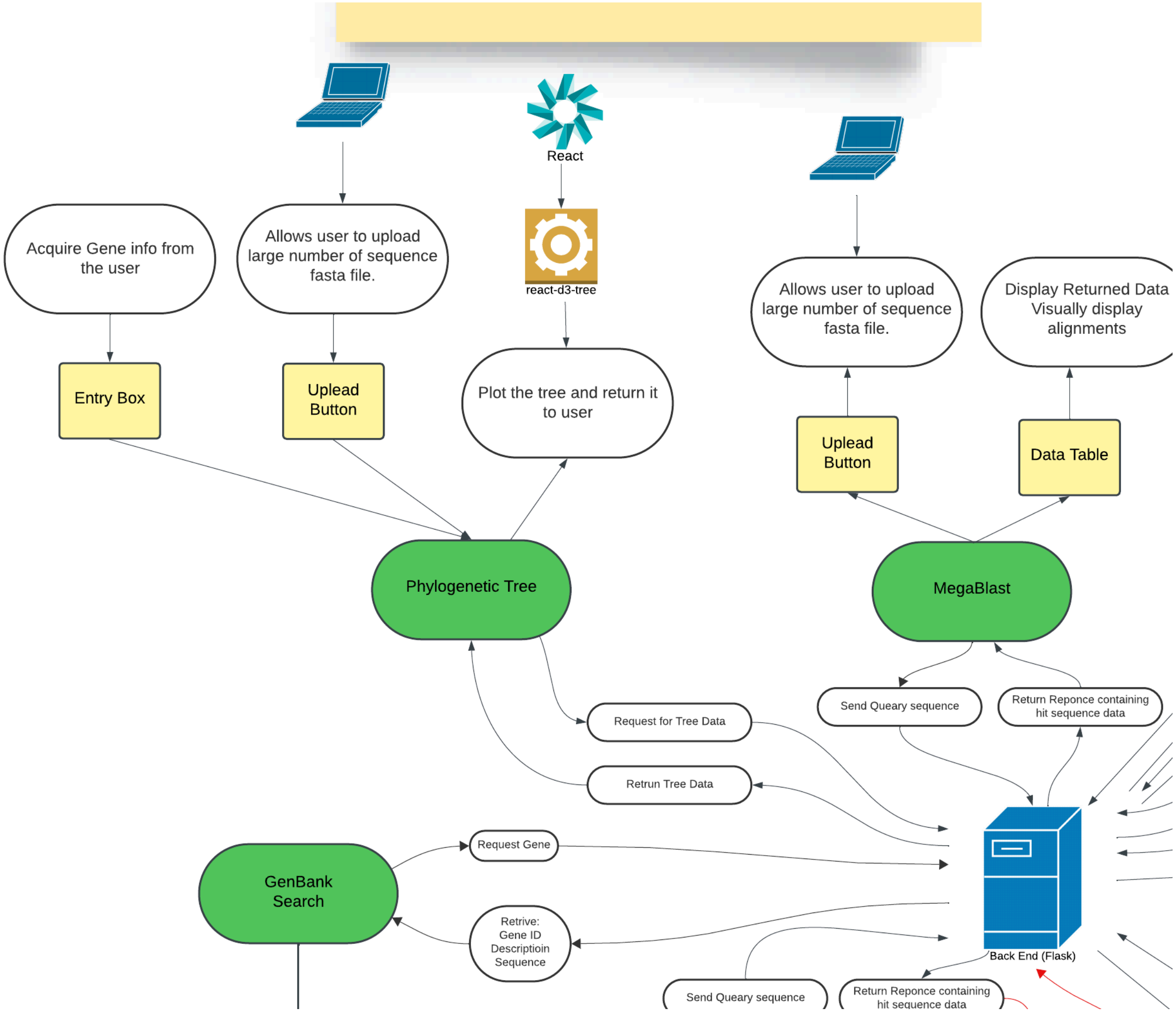

Retrun Tree Data
Request Gene
Retrive: Gene ID Descriptioin Sequence
Back End (Flask)
Send Queary sequence
Return Reponce containing hit sequence data
GearBotton
BlastNt
Save Selected Data in Fasta format
React Virtual Dom
Data Table
GearBotton
Retruning Pgylogenic Tree Image
Send Request For Alignment and Tree Creation
Display Phylogenic Tree
Cash Result for future use
Display Returned Data and make selection
Save Selected Data in Fasta format
Visually Display Selected Allignment
Retruned the fetched data to GenBank Component
Send Request From Genbank search
GenBank Route
Return XML Quary result from entraz API
Send Quary to entraz API
NCBI
Send Request Quary
Return Respons Data
Convert XML response to json
xml2json
Entraz Server
Retruning calculated rigions
Send Request For Alignment and Conservative Rigion Calculation
Display conservative regions

Note: NCBI database management system has limitations on the length of the quarry sequence, it strongly recommended for the user to USE sequence id to perform the blast alignment.


Allows user to upload large number of sequence fasta file.
Display Returned Data Visually display alignments
Upload Button
Data Table
MegaBlast
Send Queary sequence
Return Reponce containing hit sequence data
Back End (Flask)

ii. Megablast is a new improved version of the Blast. It is 10 times faster and it is very suitable for sequences that are nearly identical and differ by small percentage from another (NCBI Mega Blast documentation). MegaBLAST allows the rapid mapping of a transcript, and is useful for processing large batches of sequences (NCBI). Our megablast sequence sends the query sequence to the back end and displays the returned hit sequence. A very helpful feature of this particular component is allowing users to upload a large fasta sequence file for alignment. Additionally display alignment between query sequence and hit sequence visually.

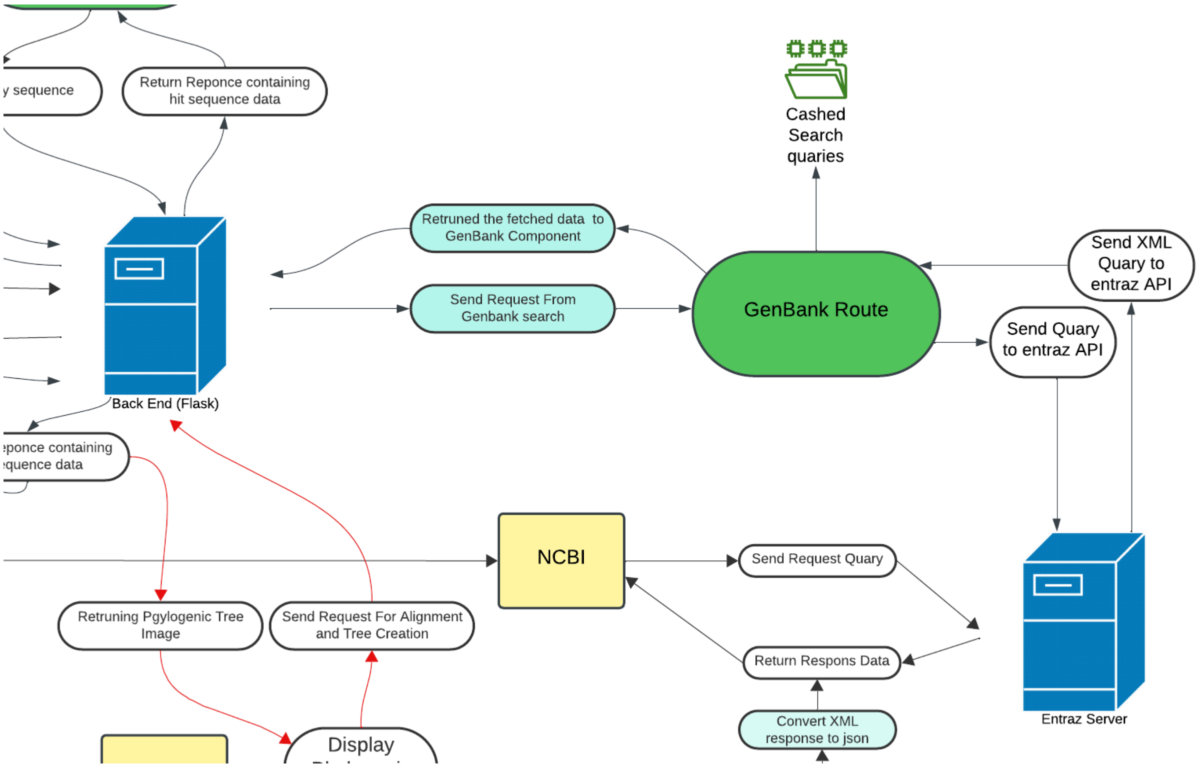
c. A dedicated phylogenetic tree generator: Gives users the ability to visualize evolutionary relationships between species by making a phylogenetic tree by simply inserting desired gene Accession IDs. This component requests data needed for constructing a tree from our back end and uses react-d3-tree to construct the tree.


React
react-d3-tree
Acquire Gene info from the user
Allows user to upload large number of sequence fasta file.
Entry Box
Uplead Button
Plot the tree and return it to user
Allows user to upload large number of sequence fasta file.
Display Returned Data Visually display alignments
Uplead Button
Data Table
Phylogenetic Tree
MegaBlast
Send Queary sequence
Return Reponce containing hit sequence data
Request for Tree Data
Retrun Tree Data
Request Gene
GenBank Search
Retrive: Gene ID Descriptioin Sequence
Back End (Flask)
Send Queary sequence
Return Reponce containing hit sequence data

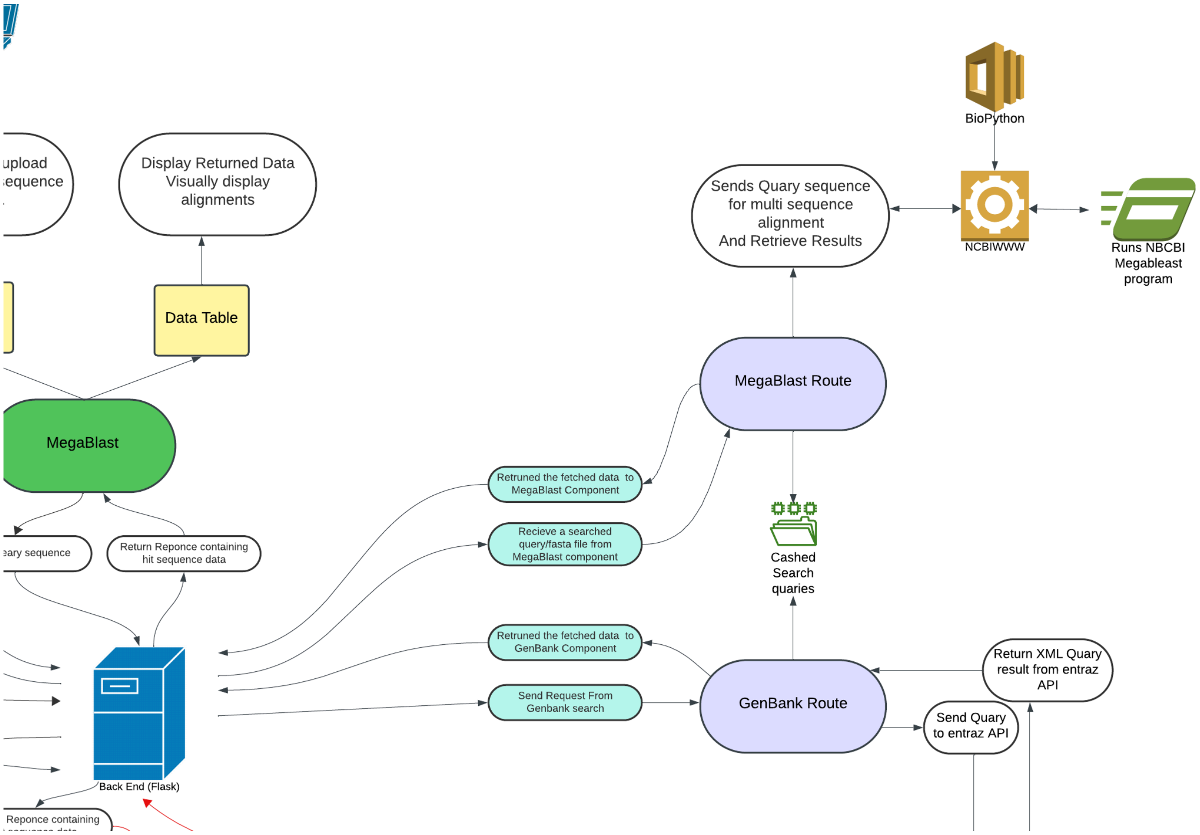
2. Backend development:

For the backend, a Flask server is implemented to create routes and endpoints for the front end to communicate with. The back end created the port 5000 that listens to the requests from the front end sent from the port 3000. It uses its components to generate responses for the front end to display. Each route created for a designated component of the front end except the blast request, that is handled from the front end.

a. GenBank search route: The /api/genbank_search is an endpoint that sends the query received from the genbank search component to the NCBI Entrez API to perform Gene search and returns the search results to the front end. A cache system was set up to store previous search results to reduce load times.

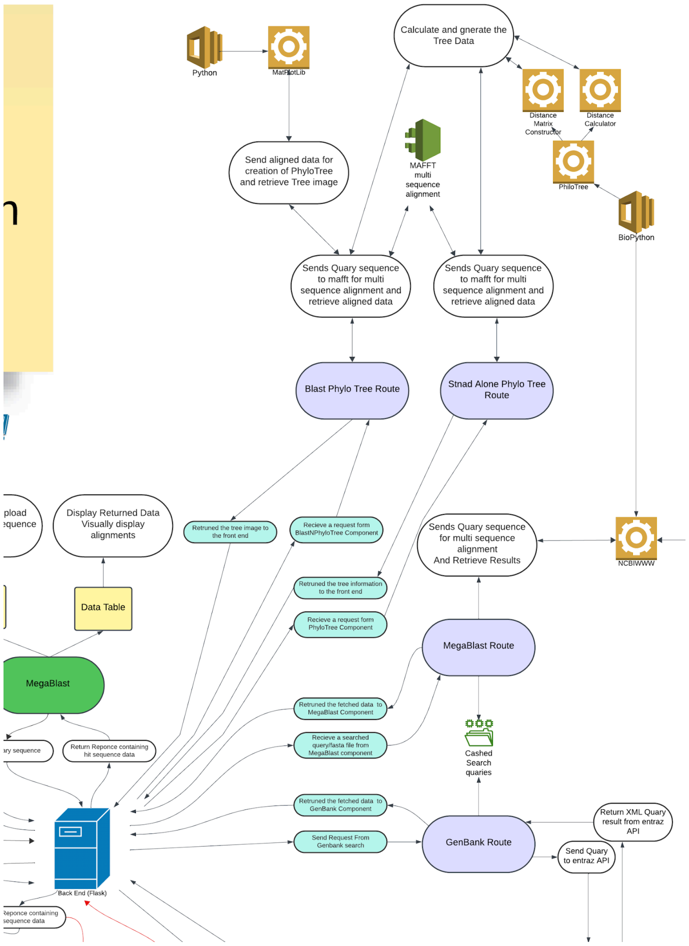

y sequence
Return Reponce containing hit sequence data
Cashed Search quaries
Retruned the fetched data to GenBank Component
Send Request From Genbank search
GenBank Route
Send XML Quary to entraz API
Send Quary to entraz API
Back End (Flask)
eponce containing equence data
NCBI
Send Request Quary
Retruning Pgylogenic Tree Image
Send Request For Alignment and Tree Creation
Return Respons Data
Convert XML response to json
Entraz Server
Display

b. MegaBLAST route: Thist endpoint uses the NCBIWWW module from BioPython to run BLAST queries from the NCBI nucleotide database. Users can paste sequences directly or upload a FASTA file.



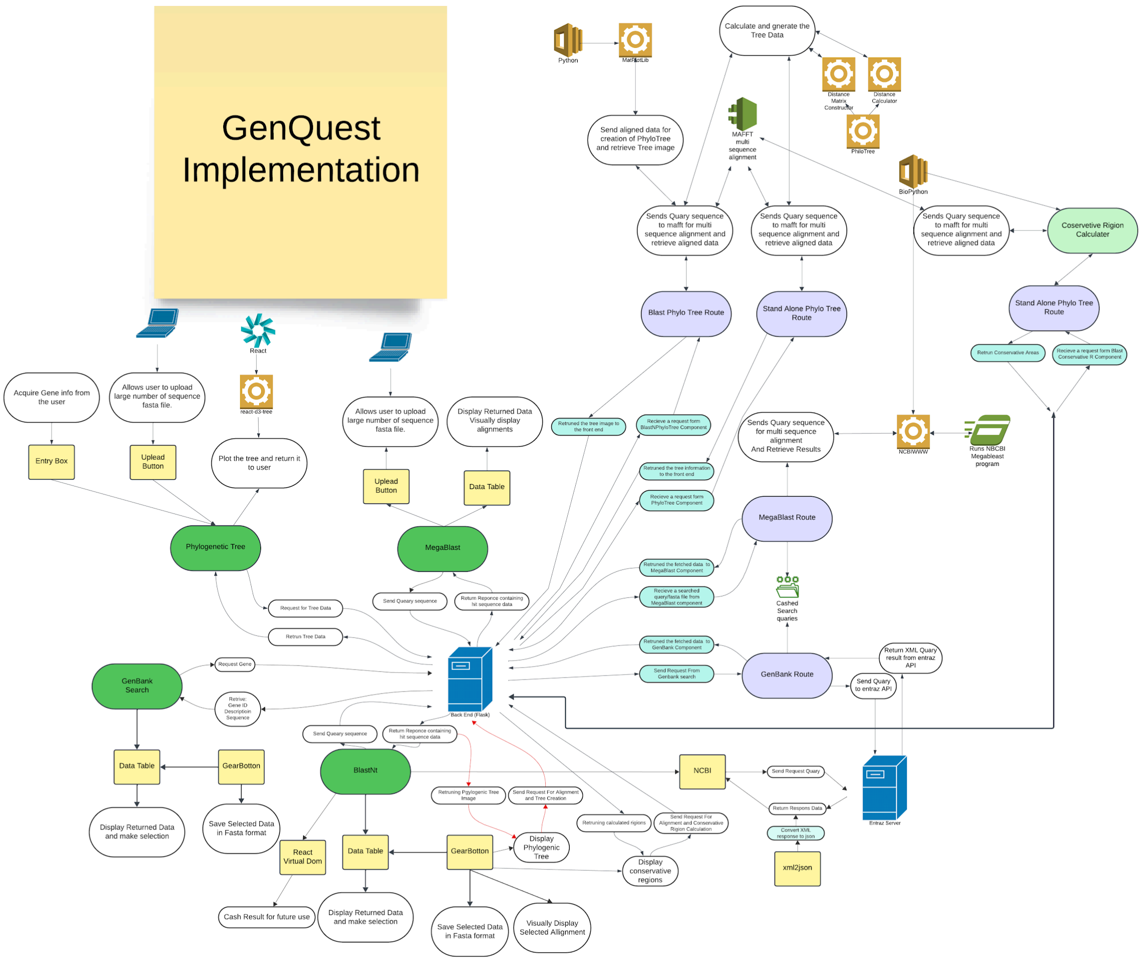
c. Stand alone Phylogenetic tree endpoint: The /api/phylo_tree provides a route for our PhyloTree front end component to perform multiple sequence alignments using MAFFT and builds the trees using UPGMA (Unweighted Pair Group Method with Averaging) method. A cache system is also implemented here.

d. Blastn Phylogenetic Tree: The /gb_dat_anl provides an endpoint for the Blastn component. It received gene ids and their relative sequences. Similarly uses MAFFT to apply multiple sequence alignments. BioPython's TreeConstruction library utilizes Distance Calculator, DistanceTreeConstructor packages to calculate distance and construct the tree. Then matplotlib uses the generated information to construct the tree and saves the result in an image file and sends that image file to the front end to be displayed.


Python
MatPlotLib
Calculate and gnerate the Tree Data
Distance Matrix Constructor
Distance Calculator
PhiloTree
BioPython
Send aligned data for creation of PhyloTree and retrieve Tree image
MAFFT multi sequence alignment
Sends Quary sequence to mafft for multi sequence alignment and retrieve aligned data
Sends Quary sequence to mafft for multi sequence alignment and retrieve aligned data
Blast Phylo Tree Route
Stnad Alone Phylo Tree Route
pload equence
Display Returned Data Visually display alignments
Retruned the tree image to the front end
Recieve a request form BlastNPhyloTree Component
Sends Quary sequence for multi sequence alignment And Retrieve Results
NCBIWWW
Retruned the tree information to the front end
Recieve a request form PhyloTree Component
Data Table
MegaBlast Route
MegaBlast
Retruned the fetched data to MegaBlast Component
Recieve a searched query/fasta file from MegaBlast component
Cashed Search quaries
ary sequence
Return Reponce containing hit sequence data
Retruned the fetched data to GenBank Component
Send Request From Genbank search
GenBank Route
Return XML Quary result from entraz API
Send Quary to entraz API
Back End (Flask)
Reponce containing sequence data

e. Conservative Region Route: The `/conservativeRegion` provides endpoints for the front end blast conservative region request. Similar to our tree generation component it sends the sequence information to MAFFT CLI (Command Line Interface) for alignment. It uses the ConservedRegionCalc.py component to calculate the conserved region and finally returns the result to the front end to display it.

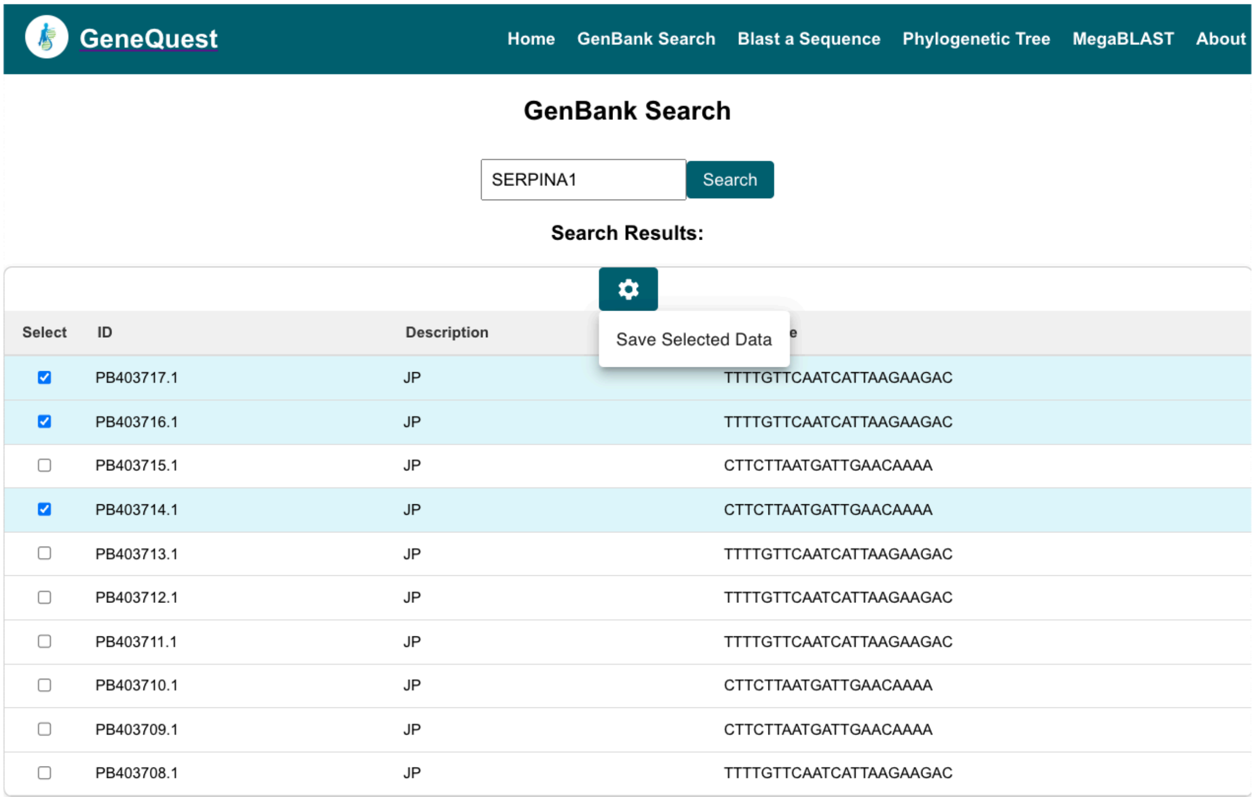
f. MAFFT (Multiple Alignment using Fast Fourier Transform): Is a software developed to perform multiple sequence alignment by applying Fast Fourier transformation. This transformation identifies homologous regions within the gene sequences, making it particularly efficient for large datasets (MAFFT documentation).


GenQuest Implementation
Python
MatPlotLib
Calculate and gnerate the Tree Data
Distance Matrix Constructor
Distance Calculator
PhiloTree
MAFFT multi sequence alignment
BioPython
Send aligned data for creation of PhyloTree and retrieve Tree image
Sends Quary sequence to mafft for multi sequence alignment and retrieve aligned data
Sends Quary sequence to mafft for multi sequence alignment and retrieve aligned data
Sends Quary sequence to mafft for multi sequence alignment and retrieve aligned data
Coservetive Rigion Calculater
Blast Phylo Tree Route
Stand Alone Phylo Tree Route
Stand Alone Phylo Tree Route
Retrun Conservative Areas
Recieve a request form Blast Conservative R Component
React
react-d3-tree
Acquire Gene info from the user
Allows user to upload large number of sequence fasta file.
Entry Box
Uplead Button
Plot the tree and return it to user
Allows user to upload large number of sequence fasta file.
Display Returned Data Visually display alignments
Uplead Button
Data Table
Retruned the tree image to the front end
Recieve a request form BlastNPhyloTree Component
Sends Quary sequence for multi sequence alignment And Retrieve Results
NCBIWWW
Runs NBCBI Megableast program
Retruned the tree information to the front end
Recieve a request form PhyloTree Component
MegaBlast Route
Phylogenetic Tree
MegaBlast
Send Queary sequence
Return Reponce containing hit sequence data
Request for Tree Data
Retrun Tree Data
Retruned the fetched data to MegaBlast Component
Recieve a searched query/fasta file from MegaBlast component
Cashed Search quaries
Retruned the fetched data to GenBank Component
Send Request From Genbank search
GenBank Route
Return XML Quary result from entraz API
Send Quary to entraz API
GenBank Search
Request Gene
Retrive: Gene ID Descriptioin Sequence
Back End (Flask)
Send Queary sequence
Return Reponce containing hit sequence data
BlastNt
NCBI
Send Request Quary
Return Respons Data
Convert XML response to json
xml2json
Entraz Server
Data Table
GearBotton
Display Returned Data and make selection
Save Selected Data in Fasta format
Retruning Pgylogenic Tree Image
Send Request For Alignment and Tree Creation
Retruning calculated rigions
Send Request For Alignment and Conservative Rigion Calculation
React Virtual Dom
Data Table
GearBotton
Display Phylogenic Tree
Display conservative regions
Cash Result for future use
Display Returned Data and make selection
Save Selected Data in Fasta format
Visually Display Selected Allignment

Results

Developing GeneQuest provided us with a fully functional web application that greatly simplifies genetic data analysis. We tested it to see how well it performed GenBank searches, BLAST/MegaBLAST alignments, implemented conservative regions calculation, and phylogenetic tree building. Here's a breakdown of the results of our web application:

1. GenBank search: For testing the genbank research result we decided to use the SERPINA1 gene. This gene provides instructions for producing alpha-1 protein that protects lungs from damages (MedLinePlus).



GenBank Search

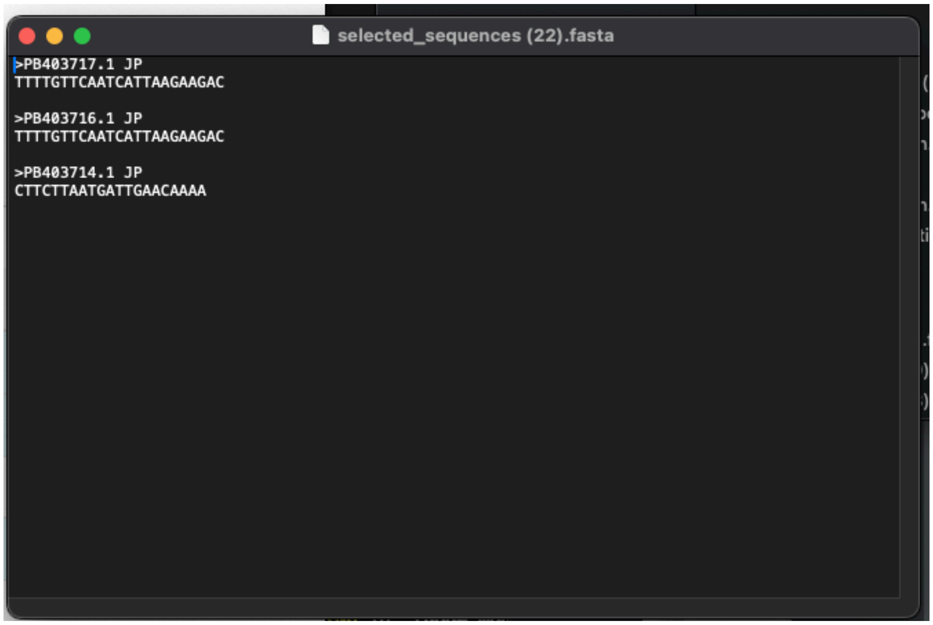
SERPINA1

Search Results:

Select	ID	Description	e
☑	PB403717.1	JP	TTTTGTTCAATCATTAAGAAGAC
☑	PB403716.1	JP	TTTTGTTCAATCATTAAGAAGAC
□	PB403715.1	JP	CTTCTTAATGATTGAACAAAA
☑	PB403714.1	JP	CTTCTTAATGATTGAACAAAA
□	PB403713.1	JP	TTTTGTTCAATCATTAAGAAGAC
□	PB403712.1	JP	TTTTGTTCAATCATTAAGAAGAC
□	PB403711.1	JP	TTTTGTTCAATCATTAAGAAGAC
□	PB403710.1	JP	CTTCTTAATGATTGAACAAAA
□	PB403709.1	JP	CTTCTTAATGATTGAACAAAA
□	PB403708.1	JP	TTTTGTTCAATCATTAAGAAGAC

Figure 1: A search for "SERPINA1" returned sequence results from NCBI GenBank, with data such as sequence ID and the DNA strands

.



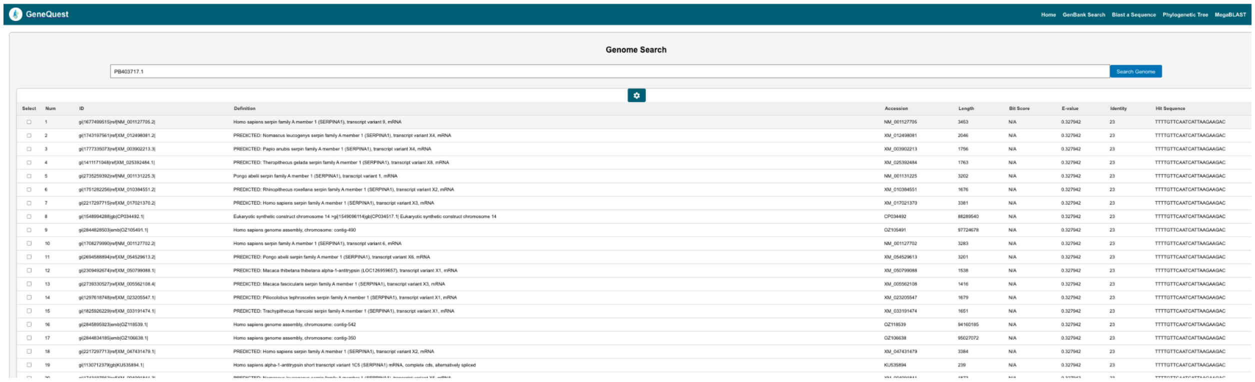
```
>PB403717.1 JP
TTTTGTTCAATCATTAAGAAGAC

>PB403716.1 JP
TTTTGTTCAATCATTAAGAAGAC

>PB403714.1 JP
CTTCTTAATGATTGAACAAAA
```

Figure 2: Result of downloaded Fasta file for SERPINA1.

2. BLAST a sequence: Following up we used PB403717.1 id and performed a blast on it. As we can see our blast returns all the hit Sequences with their id, description, e-values, hit sequence and so on.



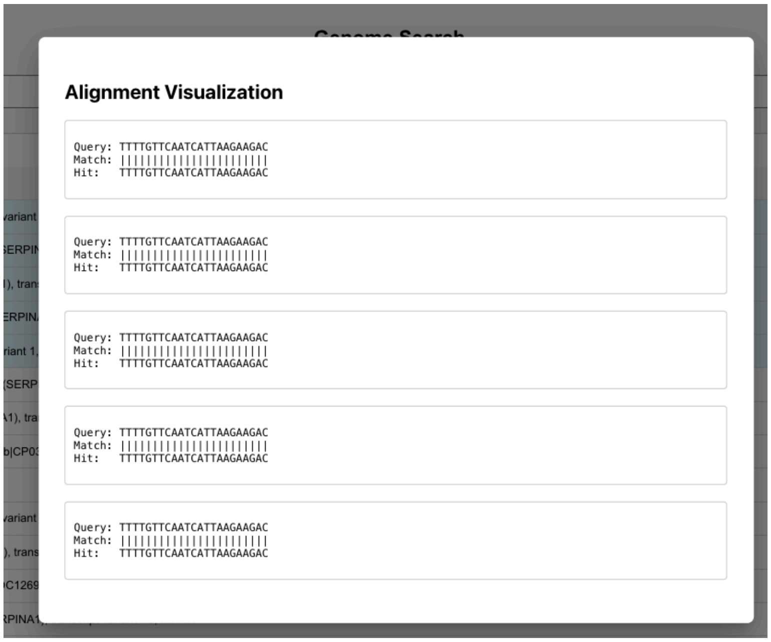
Genome Search

PB403717.1

Select	Num	ID	Definition	Accession	Length	Bit Score	E-value	Identity	Hit Sequence
□	1	gi|1677499515|ref|NM_001127705.2|	Homo sapiens serpin family A member 1 (SERPINA1), transcript variant 9, mRNA	NM_001127705	3453	N/A	0.327942	23	TTTTGTTCAATCATTAAGAAGAC
□	2	gi|1743197561|ref|XM_012498081.2|	PREDICTED: Nomascus leucogenys serpin family A member 1 (SERPINA1), transcript variant X4, mRNA	XM_012498081	2046	N/A	0.327942	23	TTTTGTTCAATCATTAAGAAGAC
□	3	gi|1777335073|ref|XM_003902213.3|	PREDICTED: Papio anubis serpin family A member 1 (SERPINA1), transcript variant X4, mRNA	XM_003902213	1756	N/A	0.327942	23	TTTTGTTCAATCATTAAGAAGAC
□	4	gi|1411171048|ref|XM_025392484.1|	PREDICTED: Theropithecus gelada serpin family A member 1 (SERPINA1), transcript variant X8, mRNA	XM_025392484	1763	N/A	0.327942	23	TTTTGTTCAATCATTAAGAAGAC
□	5	gi|2735259392|ref|NM_001131225.3|	Pongo abelii serpin family A member 1 (SERPINA1), transcript variant 1, mRNA	NM_001131225	3202	N/A	0.327942	23	TTTTGTTCAATCATTAAGAAGAC
□	6	gi|1751282256|ref|XM_010384551.2|	PREDICTED: Rhinopithecus roxellana serpin family A member 1 (SERPINA1), transcript variant X2, mRNA	XM_010384551	1676	N/A	0.327942	23	TTTTGTTCAATCATTAAGAAGAC
□	7	gi|2217297715|ref|XM_017021370.2|	PREDICTED: Homo sapiens serpin family A member 1 (SERPINA1), transcript variant X3, mRNA	XM_017021370	3381	N/A	0.327942	23	TTTTGTTCAATCATTAAGAAGAC
□	8	gi|1548994288|gb|CP034492.1|	Eukaryotic synthetic construct chromosome 14 >gi|1549096114|gb|CP034517.1| Eukaryotic synthetic construct chromosome 14	CP034492	88289540	N/A	0.327942	23	TTTTGTTCAATCATTAAGAAGAC
□	9	gi|2844828503|emb|OZ105491.1|	Homo sapiens genome assembly, chromosome: contig-490	OZ105491	97724678	N/A	0.327942	23	TTTTGTTCAATCATTAAGAAGAC
□	10	gi|1708279990|ref|NM_001127702.2|	Homo sapiens serpin family A member 1 (SERPINA1), transcript variant 6, mRNA	NM_001127702	3283	N/A	0.327942	23	TTTTGTTCAATCATTAAGAAGAC
□	11	gi|2694588894|ref|XM_054529613.2|	PREDICTED: Pongo abelii serpin family A member 1 (SERPINA1), transcript variant X6, mRNA	XM_054529613	3201	N/A	0.327942	23	TTTTGTTCAATCATTAAGAAGAC
□	12	gi|2309492674|ref|XM_050799088.1|	PREDICTED: Macaca thibetana thibetana alpha-1-antitrypsin (LOC126959657), transcript variant X1, mRNA	XM_050799088	1538	N/A	0.327942	23	TTTTGTTCAATCATTAAGAAGAC
□	13	gi|2739330527|ref|XM_005562108.4|	PREDICTED: Macaca fascicularis serpin family A member 1 (SERPINA1), transcript variant X3, mRNA	XM_005562108	1416	N/A	0.327942	23	TTTTGTTCAATCATTAAGAAGAC
□	14	gi|1297618748|ref|XM_023205547.1|	PREDICTED: Piliocolobus tephrosceles serpin family A member 1 (SERPINA1), transcript variant X1, mRNA	XM_023205547	1679	N/A	0.327942	23	TTTTGTTCAATCATTAAGAAGAC
□	15	gi|1825926229|ref|XM_033191474.1|	PREDICTED: Trachypithecus francoisi serpin family A member 1 (SERPINA1), transcript variant X1, mRNA	XM_033191474	1651	N/A	0.327942	23	TTTTGTTCAATCATTAAGAAGAC
□	16	gi|2845895923|emb|OZ118539.1|	Homo sapiens genome assembly, chromosome: contig-542	OZ118539	94160185	N/A	0.327942	23	TTTTGTTCAATCATTAAGAAGAC
□	17	gi|2844834185|emb|OZ106638.1|	Homo sapiens genome assembly, chromosome: contig-350	OZ106638	95027072	N/A	0.327942	23	TTTTGTTCAATCATTAAGAAGAC
□	18	gi|2217297713|ref|XM_047431479.1|	PREDICTED: Homo sapiens serpin family A member 1 (SERPINA1), transcript variant X2, mRNA	XM_047431479	3384	N/A	0.327942	23	TTTTGTTCAATCATTAAGAAGAC
□	19	gi|1130712379|gb|KU535894.1|	Homo sapiens alpha-1-antitrypsin short transcript variant 1C5 (SERPINA1) mRNA, complete cds, alternatively spliced	KU535894	239	N/A	0.327942	23	TTTTGTTCAATCATTAAGAAGAC
□	20	gi|1743197563|ref|XM_004091841.3|	PREDICTED: Nomascus leucogenys serpin family A member 1 (SERPINA1), transcript variant X5, mRNA	XM_004091841	1873	N/A	0.327942	23	TTTTGTTCAATCATTAAGAAGAC

Figure 2.1: A user entered search query returned BLAST results from NCBI's Blast server

a. After the results were displayed we selected the top five results from the table. There is a 'Gear' icon which enables the user to do operations like show alignment, generate a phylogenetic tree, calculate the conservative region and save the data.



Alignment Visualization

```
Query: TTTTGTTCAATCATTAAGAAGAC
Match: |||||||||||||||||||||||
Hit:   TTTTGTTCAATCATTAAGAAGAC
```

```
Query: TTTTGTTCAATCATTAAGAAGAC
Match: |||||||||||||||||||||||
Hit:   TTTTGTTCAATCATTAAGAAGAC
```

```
Query: TTTTGTTCAATCATTAAGAAGAC
Match: |||||||||||||||||||||||
Hit:   TTTTGTTCAATCATTAAGAAGAC
```

```
Query: TTTTGTTCAATCATTAAGAAGAC
Match: |||||||||||||||||||||||
Hit:   TTTTGTTCAATCATTAAGAAGAC
```

```
Query: TTTTGTTCAATCATTAAGAAGAC
Match: |||||||||||||||||||||||
Hit:   TTTTGTTCAATCATTAAGAAGAC
```

Figure 2.2: Alignment visualization are shown after users select desired results, clicks in the gear icon and selects show alignment

b. For testing out the phylogenetic tree I used these accession number for chimpanzee dna accession number PB145443.1, then we generated a tree from 5 closest one and this is the resulting phylogenetic tree.

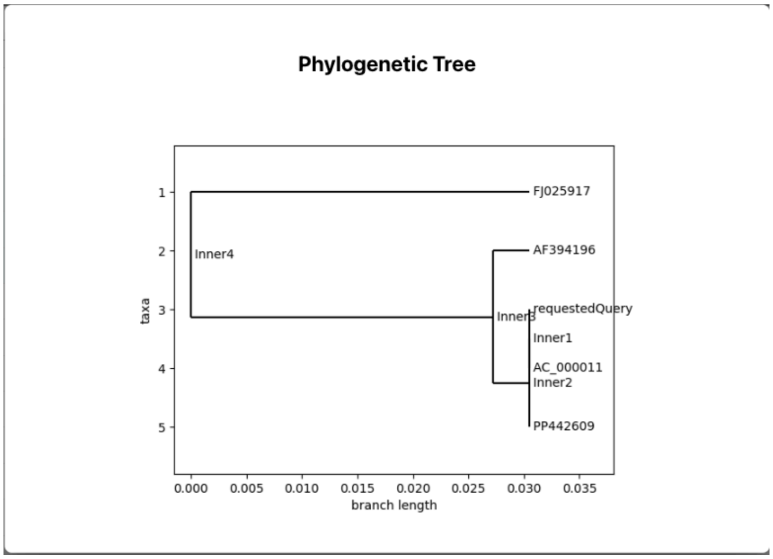
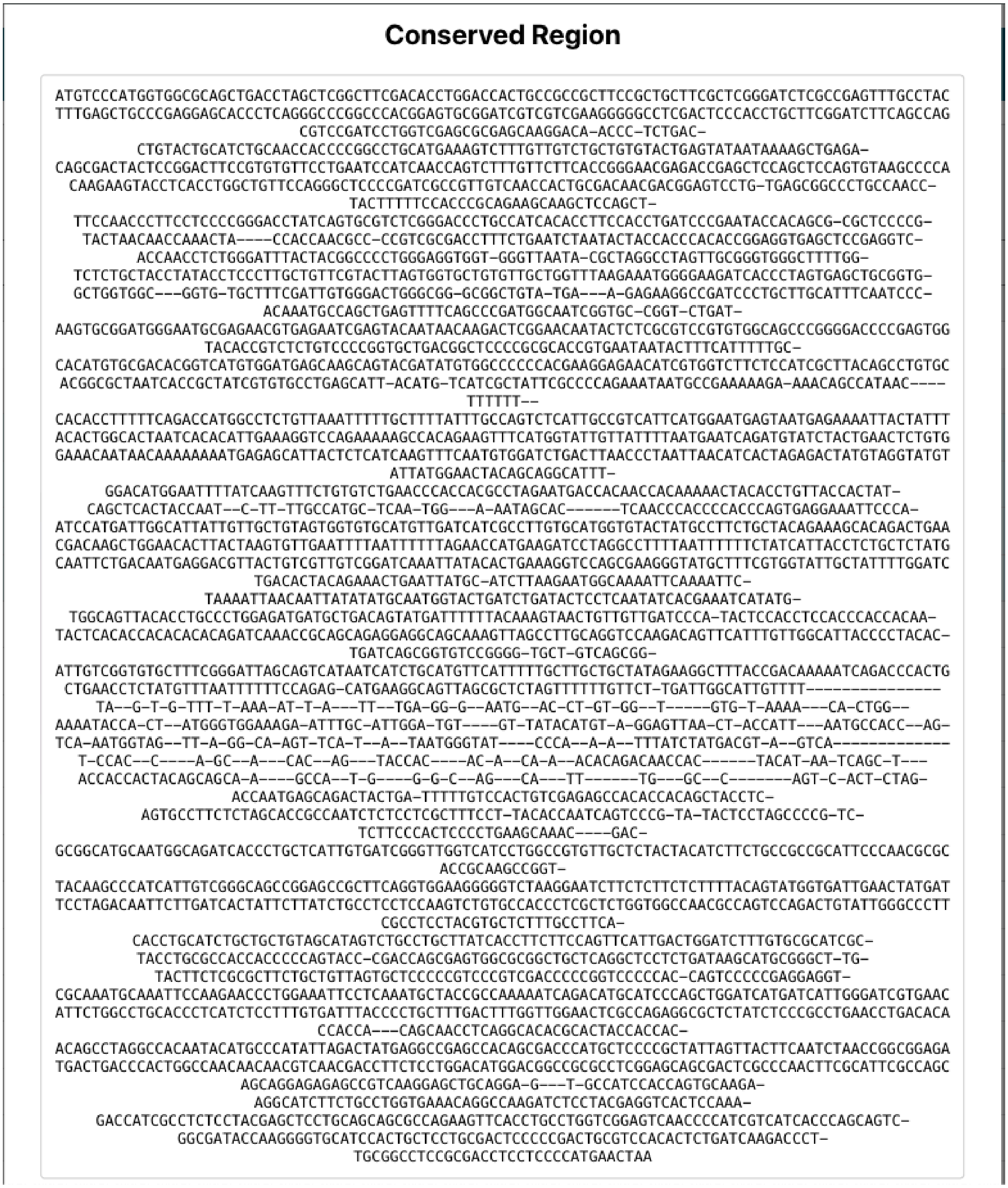


Figure 2.3: Phylogenetic tree generated as an image and shown on the frontend directly on the “Blast a sequence” page

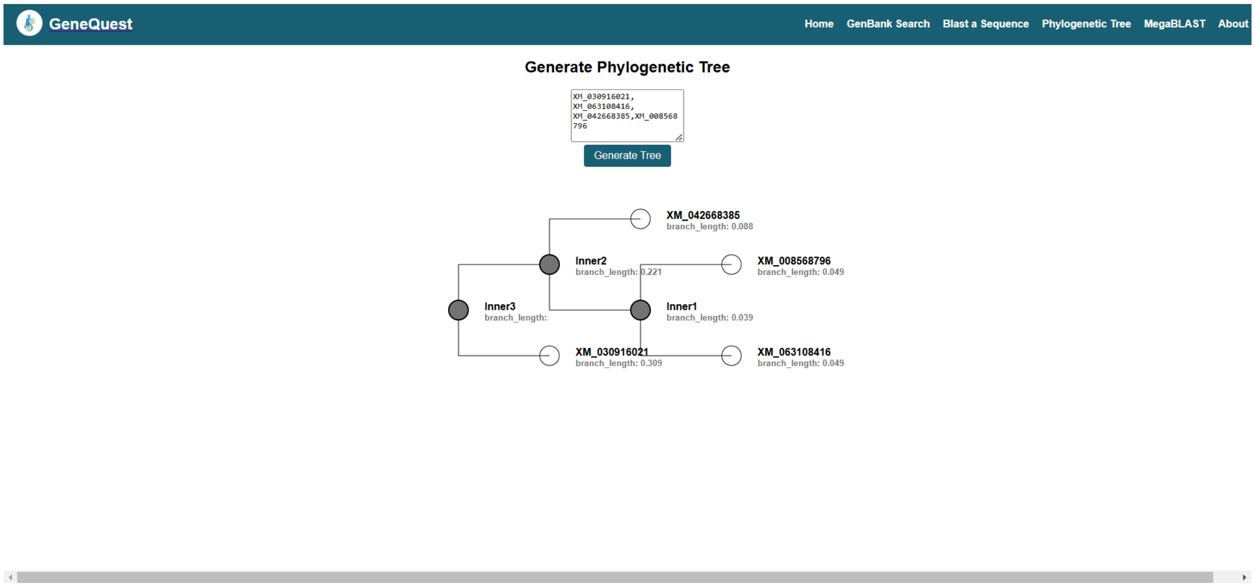
c. Follow up this is the conservative region of the selected sequences.



Conserved Region

```
ATGTCCCATGGTGGCGCAGCTGACCTAGCTCGGCTTCGACACCTGGACCACTGCCGCCGCTTCCGCTGCTTCGCTCGGGATCTCGCCGAGTTTGCCTAC
TTTGAGCTGCCCGAGGAGCACCCTCAGGGCCCGGCCCACGGAGTGCGGATCGTCGTCGAAGGGGGCCTCGACTCCCACCTGCTTCGGATCTTCAGCCAG
CGTCCGATCCTGGTCGAGCGCGAGCAAGGACA-ACCC-TCTGAC-
CTGTACTGCATCTGCAACCACCCCGGCCTGCATGAAAGTCTTTGTTGTCTGCTGTGTACTGAGTATAATAAAAGCTGAGA-
CAGCGACTACTCCGGACTTCCGTGTGTTCCTGAATCCATCAACCAGTCTTTGTTCTTCACCGGGAACGAGACCGAGCTCCAGCTCCAGTGTAAGCCCCA
CAAGAAGTACCTCACCTGGCTGTTCCAGGGCTCCCCGATCGCCGTTGTCAACCACTGCGACAACGACGGAGTCCTG-TGAGCGGCCCTGCCAACC-
TACTTTTTCCACCCGCAGAAGCAAGCTCCAGCT-
TTCCAACCCTTCCTCCCCGGGACCTATCAGTGCGTCTCGGGACCCTGCCATCACACCTTCCACCTGATCCCGAATACCACAGCG-CGCTCCCCG-
TACTAACAACCAAACTA----CCACCAACGCC-CCGTCGCGACCTTTCTGAATCTAATACTACCACCCACACCGGAGGTGAGCTCCGAGGTC-
ACCAACCTCTGGGATTTACTACGGCCCCTGGGAGGTGGT-GGGTTAATA-CGCTAGGCCTAGTTGCGGGTGGGCTTTTGG-
TCTCTGCTACCTATACCTCCCTTGCTGTTCGTACTTAGTGGTGCTGTGTTGCTGGTTTAAGAAATGGGGAAGATCACCCTAGTGAGCTGCGGTG-
GCTGGTGGC---GGTG-TGCTTTCGATTGTGGGACTGGGCGG-GCGGCTGTA-TGA---A-GAGAAGGCCGATCCCTGCTTGCATTTCAATCCC-
ACAAATGCCAGCTGAGTTTTCAGCCCGATGGCAATCGGTGC-CGGT-CTGAT-
AAGTGCGGATGGGAATGCGAGAACGTGAGAATCGAGTACAATAACAAGACTCGGAACAATACTCTCGCGTCCGTGTGGCAGCCCGGGGACCCCGAGTGG
TACACCGTCTCTGTCCCCGGTGCTGACGGCTCCCCGCGCACCGTGAATAATACTTTCATTTTTGC-
CACATGTGCGACACGGTCATGTGGATGAGCAAGCAGTACGATATGTGGCCCCCCACGAAGGAGAACATCGTGGTCTTCTCCATCGCTTACAGCCTGTGC
ACGGCGCTAATCACCGCTATCGTGTGCCTGAGCATT-ACATG-TCATCGCTATTCGCCCCAGAAATAATGCCGAAAAAGA-AAACAGCCATAAC----
TTTTTT--
CACACCTTTTTCAGACCATGGCCTCTGTTAAATTTTTGCTTTTATTTGCCAGTCTCATTGCCGTCATTCATGGAATGAGTAATGAGAAAATTACTATTT
ACACTGGCACTAATCACACATTGAAAGGTCCAGAAAAAGCCACAGAAGTTTCATGGTATTGTTATTTTAATGAATCAGATGTATCTACTGAACTCTGTG
GAAACAATAACAAAAAAAATGAGAGCATTACTCTCATCAAGTTTCAATGTGGATCTGACTTAACCCTAATTAACATCACTAGAGACTATGTAGGTATGT
ATTATGGAACTACAGCAGGCATTT-
GGACATGGAATTTTATCAAGTTTCTGTGTCTGAACCCACCACGCCTAGAATGACCACAACCACAAAAACTACACCTGTTACCACTAT-
CAGCTCACTACCAAT--C-TT-TTGCCATGC-TCAA-TGG---A-AATAGCAC------TCAACCCACCCCACCCAGTGAGGAAATTCCCA-
ATCCATGATTGGCATTATTGTTGCTGTAGTGGTGTGCATGTTGATCATCGCCTTGTGCATGGTGTACTATGCCTTCTGCTACAGAAAGCACAGACTGAA
CGACAAGCTGGAACACTTACTAAGTGTTGAATTTTAATTTTTTAGAACCATGAAGATCCTAGGCCTTTTAATTTTTTCTATCATTACCTCTGCTCTATG
CAATTCTGACAATGAGGACGTTACTGTCGTTGTCGGATCAAATTATACACTGAAAGGTCCAGCGAAGGGTATGCTTTCGTGGTATTGCTATTTTGGATC
TGACACTACAGAAACTGAATTATGC-ATCTTAAGAATGGCAAAATTCAAAATTC-
TAAAATTAACAATTATATATGCAATGGTACTGATCTGATACTCCTCAATATCACGAAATCATATG-
TGGCAGTTACACCTGCCCTGGAGATGATGCTGACAGTATGATTTTTTACAAAGTAACTGTTGTTGATCCCA-TACTCCACCTCCACCCACCACAA-
TACTCACACCACACACACAGATCAAACCGCAGCAGAGGAGGCAGCAAAGTTAGCCTTGCAGGTCCAAGACAGTTCATTTGTTGGCATTACCCCTACAC-
TGATCAGCGGTGTCCGGGG-TGCT-GTCAGCGG-
ATTGTCGGTGTGCTTTCGGGATTAGCAGTCATAATCATCTGCATGTTCATTTTTGCTTGCTGCTATAGAAGGCTTTACCGACAAAAATCAGACCCACTG
CTGAACCTCTATGTTTAATTTTTTCCAGAG-CATGAAGGCAGTTAGCGCTCTAGTTTTTTGTTCT-TGATTGGCATTGTTTT---------------
TA--G-T-G-TTT-T-AAA-AT-T-A---TT--TGA-GG-G--AATG--AC-CT-GT-GG--T-----GTG-T-AAAA---CA-CTGG--
AAAATACCA-CT--ATGGGTGGAAAGA-ATTTGC-ATTGGA-TGT----GT-TATACATGT-A-GGAGTTAA-CT-ACCATT---AATGCCACC--AG-
TCA-AATGGTAG--TT-A-GG-CA-AGT-TCA-T--A--TAATGGGTAT----CCCA--A-A--TTTATCTATGACGT-A--GTCA-------------
T-CCAC--C----A-GC--A---CAC--AG---TACCAC----AC-A--CA-A--ACACAGACAACCAC------TACAT-AA-TCAGC-T---
ACCACCACTACAGCAGCA-A----GCCA--T-G----G-G-C--AG---CA---TT------TG---GC--C-------AGT-C-ACT-CTAG-
ACCAATGAGCAGACTACTGA-TTTTTGTCCACTGTCGAGAGCCACACCACAGCTACCTC-
AGTGCCTTCTCTAGCACCGCCAATCTCTCCTCGCTTTCCT-TACACCAATCAGTCCCG-TA-TACTCCTAGCCCCG-TC-
TCTTCCCACTCCCCTGAAGCAAAC----GAC-
GCGGCATGCAATGGCAGATCACCCTGCTCATTGTGATCGGGTTGGTCATCCTGGCCGTGTTGCTCTACTACATCTTCTGCCGCCGCATTCCCAACGCGC
ACCGCAAGCCGGT-
TACAAGCCCATCATTGTCGGGCAGCCGGAGCCGCTTCAGGTGGAAGGGGGTCTAAGGAATCTTCTCTTCTCTTTTACAGTATGGTGATTGAACTATGAT
TCCTAGACAATTCTTGATCACTATTCTTATCTGCCTCCTCCAAGTCTGTGCCACCCTCGCTCTGGTGGCCAACGCCAGTCCAGACTGTATTGGGCCCTT
CGCCTCCTACGTGCTCTTTGCCTTCA-
CACCTGCATCTGCTGCTGTAGCATAGTCTGCCTGCTTATCACCTTCTTCCAGTTCATTGACTGGATCTTTGTGCGCATCGC-
TACCTGCGCCACCACCCCCAGTACC-CGACCAGCGAGTGGCGCGGCTGCTCAGGCTCCTCTGATAAGCATGCGGGCT-TG-
TACTTCTCGCGCTTCTGCTGTTAGTGCTCCCCCGTCCCGTCGACCCCCGGTCCCCCAC-CAGTCCCCCGAGGAGGT-
CGCAAATGCAAATTCCAAGAACCCTGGAAATTCCTCAAATGCTACCGCCAAAAATCAGACATGCATCCCAGCTGGATCATGATCATTGGGATCGTGAAC
ATTCTGGCCTGCACCCTCATCTCCTTTGTGATTTACCCCTGCTTTGACTTTGGTTGGAACTCGCCAGAGGCGCTCTATCTCCCGCCTGAACCTGACACA
CCACCA---CAGCAACCTCAGGCACACGCACTACCACCAC-
ACAGCCTAGGCCACAATACATGCCCATATTAGACTATGAGGCCGAGCCACAGCGACCCATGCTCCCCGCTATTAGTTACTTCAATCTAACCGGCGGAGA
TGACTGACCCACTGGCCAACAACAACGTCAACGACCTTCTCCTGGACATGGACGGCCGCGCCTCGGAGCAGCGACTCGCCCAACTTCGCATTCGCCAGC
AGCAGGAGAGAGCCGTCAAGGAGCTGCAGGA-G---T-GCCATCCACCAGTGCAAGA-
AGGCATCTTCTGCCTGGTGAAACAGGCCAAGATCTCCTACGAGGTCACTCCAAA-
GACCATCGCCTCTCCTACGAGCTCCTGCAGCAGCGCCAGAAGTTCACCTGCCTGGTCGGAGTCAACCCCATCGTCATCACCCAGCAGTC-
GGCGATACCAAGGGGTGCATCCACTGCTCCTGCGACTCCCCCGACTGCGTCCACACTCTGATCAAGACCCT-
TGCGGCCTCCGCGACCTCCTCCCCATGAACTAA
```

3. Phylogenetic tree generation: To test the phylogenetic tree generation tool we used these accession numbers to generate the tree: XM_030916021, XM_063108416, XM_042668385,XM_008568796. These accession numbers are identifiers for Golden snub nosed monkey, Philippine flying lemur, banner tailed kangaroo rat and sunda flying lemur genome respectively. From the generated phylogenetic tree we can see that the Philippine flying lemur and the Sunda flying lemur are closer together. Thus they are evolutionarily closely related. The banner-tailed kangaroo rat is farther from the two lemur species. And the golden snub monkey is the farthest node which signifies that it has greater evolutionary distance from the other species. Thus this phylogenetic tree is used to show the evolutionary relationship between these species.



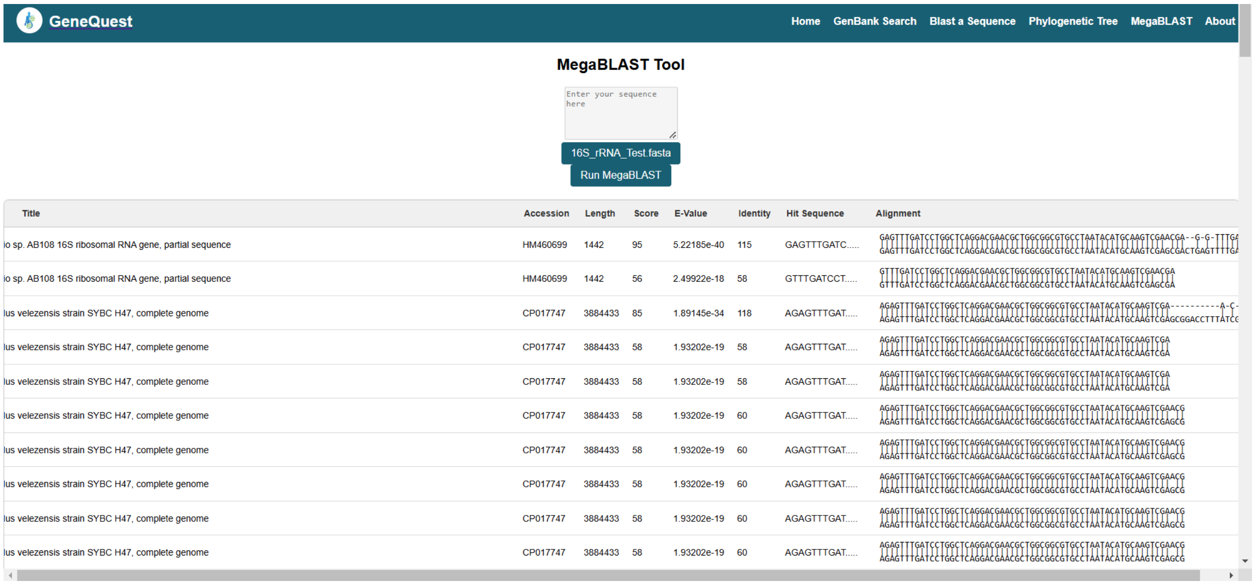
Generate Phylogenetic Tree

```
XM_030916021,
XM_063108416,
XM_042668385,XM_008568
796
```

Figure 3: A user entered accession numbers separated by commas. GeneQuest then generated a phylogenetic tree and displayed it in the frontend

4. MegaBLAST tool: To test the MegaBLAST tool we uploaded a fasta file containing a snippet of 16S rRNA gene sequence. The 16S rRNA gene is a highly conserved gene found in the ribosome of all bacteria and archaea and often used for DNA barcoding. This sequence is useful to determine relationships between microbial species. As we can see in the figure below, the

MegaBLAST tool returns a result table that contains accession numbers, lengths, scores, E-values, identity, hit sequence and alignments.



Title	Accession	Length	Score	E-Value	Identity	Hit Sequence	Alignment
io sp. AB108 16S ribosomal RNA gene, partial sequence	HM460699	1442	95	5.22185e-40	115	GAGTTTGATC.....	GAGTTTGATCCTGGCTCAGGACGAACGCTGGCGGCGTGCCTAATACATGCAAGTCGAACGA--G-G-TTTGA GAGTTTGATCCTGGCTCAGGACGAACGCTGGCGGCGTGCCTAATACATGCAAGTCGAGCGACTGAGTTTTGA
io sp. AB108 16S ribosomal RNA gene, partial sequence	HM460699	1442	56	2.49922e-18	58	GTTTGATCCT.....	GTTTGATCCTGGCTCAGGACGAACGCTGGCGGCGTGCCTAATACATGCAAGTCGAACGA GTTTGATCCTGGCTCAGGACGAACGCTGGCGGCGTGCCTAATACATGCAAGTCGAGCGA
lus velezensis strain SYBC H47, complete genome	CP017747	3884433	85	1.89145e-34	118	AGAGTTTGAT.....	AGAGTTTGATCCTGGCTCAGGACGAACGCTGGCGGCGTGCCTAATACATGCAAGTCGA----------A-C- AGAGTTTGATCCTGGCTCAGGACGAACGCTGGCGGCGTGCCTAATACATGCAAGTCGAGCGGACCTTTATCG
lus velezensis strain SYBC H47, complete genome	CP017747	3884433	58	1.93202e-19	58	AGAGTTTGAT.....	AGAGTTTGATCCTGGCTCAGGACGAACGCTGGCGGCGTGCCTAATACATGCAAGTCGA AGAGTTTGATCCTGGCTCAGGACGAACGCTGGCGGCGTGCCTAATACATGCAAGTCGA
lus velezensis strain SYBC H47, complete genome	CP017747	3884433	58	1.93202e-19	58	AGAGTTTGAT.....	AGAGTTTGATCCTGGCTCAGGACGAACGCTGGCGGCGTGCCTAATACATGCAAGTCGA AGAGTTTGATCCTGGCTCAGGACGAACGCTGGCGGCGTGCCTAATACATGCAAGTCGA
lus velezensis strain SYBC H47, complete genome	CP017747	3884433	58	1.93202e-19	60	AGAGTTTGAT.....	AGAGTTTGATCCTGGCTCAGGACGAACGCTGGCGGCGTGCCTAATACATGCAAGTCGAACG AGAGTTTGATCCTGGCTCAGGACGAACGCTGGCGGCGTGCCTAATACATGCAAGTCGAGCG
lus velezensis strain SYBC H47, complete genome	CP017747	3884433	58	1.93202e-19	60	AGAGTTTGAT.....	AGAGTTTGATCCTGGCTCAGGACGAACGCTGGCGGCGTGCCTAATACATGCAAGTCGAACG AGAGTTTGATCCTGGCTCAGGACGAACGCTGGCGGCGTGCCTAATACATGCAAGTCGAGCG
lus velezensis strain SYBC H47, complete genome	CP017747	3884433	58	1.93202e-19	60	AGAGTTTGAT.....	AGAGTTTGATCCTGGCTCAGGACGAACGCTGGCGGCGTGCCTAATACATGCAAGTCGAACG AGAGTTTGATCCTGGCTCAGGACGAACGCTGGCGGCGTGCCTAATACATGCAAGTCGAGCG
lus velezensis strain SYBC H47, complete genome	CP017747	3884433	58	1.93202e-19	60	AGAGTTTGAT.....	AGAGTTTGATCCTGGCTCAGGACGAACGCTGGCGGCGTGCCTAATACATGCAAGTCGAACG AGAGTTTGATCCTGGCTCAGGACGAACGCTGGCGGCGTGCCTAATACATGCAAGTCGAGCG
lus velezensis strain SYBC H47, complete genome	CP017747	3884433	58	1.93202e-19	60	AGAGTTTGAT.....	AGAGTTTGATCCTGGCTCAGGACGAACGCTGGCGGCGTGCCTAATACATGCAAGTCGAACG AGAGTTTGATCCTGGCTCAGGACGAACGCTGGCGGCGTGCCTAATACATGCAAGTCGAGCG

Figure 4: User uploaded a fasta file to the MegaBLAST tool. The results are show in a table format

Discussion

1. Summary:

    GeneQuest successfully integrates multiple bioinformatics tools like GenBank search, BLAST/MegaBLAST search, sequence alignments, phylogenetic tree generation and calculating conservative regions into one user-friendly web application. Our app is designed to streamline bioinformatics workflow which with the current systems can be confusing and can fragment the workflow and be time consuming. By integrating the tools mentioned above we created a web application that is user friendly and effective for genetic sequence analysis.

2. Interpretation of findings:

    a. GenBank search: Our tool simplifies searching genetic sequences. The performance of the app is improved with the implementation of a caching system.

   b. BLAST and MegaBLAST: GeneQuest supports both short and long sequence alignments by providing the users with options to use either BLAST or MegaBLAST. Which makes our application more versatile for genetic analysis.
   c. Phylogenetic tree generator: The phylogenetic tree building tool gives the user the ability to visualize evolutionary relationships and understand genetic similarities and differences in an easy and meaningful way.
   d. Conservative Region Calculator: Calculates the conservative regions of selected gene sequences for further analysis.
3. Improvements:

   One of the main limitations of GeneQuest is its reliance on other APIs. This limits us to the timing that is required to send the request and wait for the results. For example if we are searching a sequence that is new to our application, since it is not cached we have to wait on average 5 minutes for the response from NCBI APIs. A major improvement could be implementing local algorithms that reduce our reliance on slow APIs.
4. Expanding our work given unlimited resources:

   If we had unlimited resources we would like to expand GeneQuest so it is able to handle protein sequences. We would also like to implement advanced visualization for the phylogenetic tree and make the visualization interactive. We would also add educational features like tutorials to help beginners understand the bioinformatics tools GeneQuest offers and also help them to be familiar with genetic analysis and basic bioinformatic workflows.

Conclusion

GeneQuest makes bioinformatics tools more accessible to users and makes them more user friendly than the current solutions while keeping their functionalities. The integration of GenBank search, BLAST and MegaBLAST tools, phylogenetic tree generation and conservative region calculation into one single application streamlines genetic analysis workflows. GeneQuest has the potential to become a great resource for educational and professional uses.

References

Suzuki, David T., et al. *An introduction to genetic analysis*. No. Ed. 3. 1986.

a. https://www.cabidigitallibrary.org/doi/full/10.5555/19870103297

Chowdhury, B., & Garai, G. (2017). A review on multiple sequence alignment from the perspective of genetic algorithm. *Genomics*, *109*(5), 419–431. https://doi.org/10.1016/j.ygeno.2017.06.007

Maslov, D. A., Kolesnikov, A. A., & Zaitseva, G. N. (1984). Conservative and divergent base sequence regions in the maxicircle kinetoplast DNA of several trypanosomatid flagellates. *Molecular and Biochemical Parasitology*, *12*(3), 351–364. https://doi.org/10.1016/0166-6851(84)90091-4

Wheeler, D., & Bhagwat, M. (2007). BLAST QuickStart. In *Comparative Genomics: Volumes 1 and 2*. Humana Press. https://www.ncbi.nlm.nih.gov/books/NBK1734/