Environmental Sample Classification

Joshua Katz, Kurt Zimmer

Presentation Outline

- Describe the biological background
- Talk about our proposed solution
- Discuss technologies and tools we created and used to accomplish the solution
- Go over a sample usage of the website
- Show outputted data
- Compare original goals vs. what we accomplished this semester

The Problem

- Discovery of novel viruses by classifying a multitude of genetic information in environmental samples (Metagenomics)
- Go from a string of letters (‘A’, ‘T’, ‘C’, ‘G’) to an assembled genome and/or identification of the origin of the species
- Need to make use of existing biological tools and databases to turn the string of letters into meaningful information
  - The process of using these tools must be streamlined and simple so that every member of the lab can use them and save their tools and inputs for multiple sequence runs.

Proposed Solution

- Create a website that:
  - Contains useful tools to identify species from metagenomic data
  - Easy to use by anyone with a basic biological background
  - Contains a 'workflow' interface where users can save their preferred sequence of tools and their inputs to run multiple sequence files on

Tools & Technologies Pt. 1

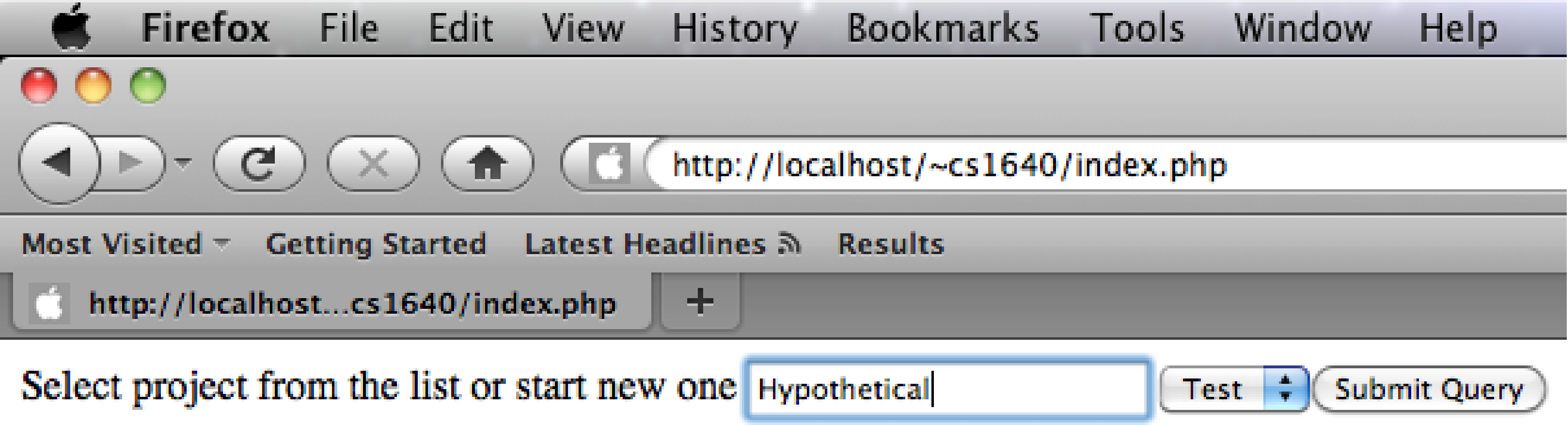
- Used these programming and scripting languages on an apache server:
  - Perl (BioPerl, GD Graphics Library)
  - PHP
  - HTML/CSS

Tools and Technologies Pt. 2

- Used existing databases and tools:
  - BLAST tool and database
  - clustalw tool
  - CD-HIT tool
  - NCBI taxonomic database
- Created these tools:
  - Project login system
  - File uploader and viewer
  - Workflow system and workflow executor
  - Wrappers for every database and tool to allow the user to use them from a website interface

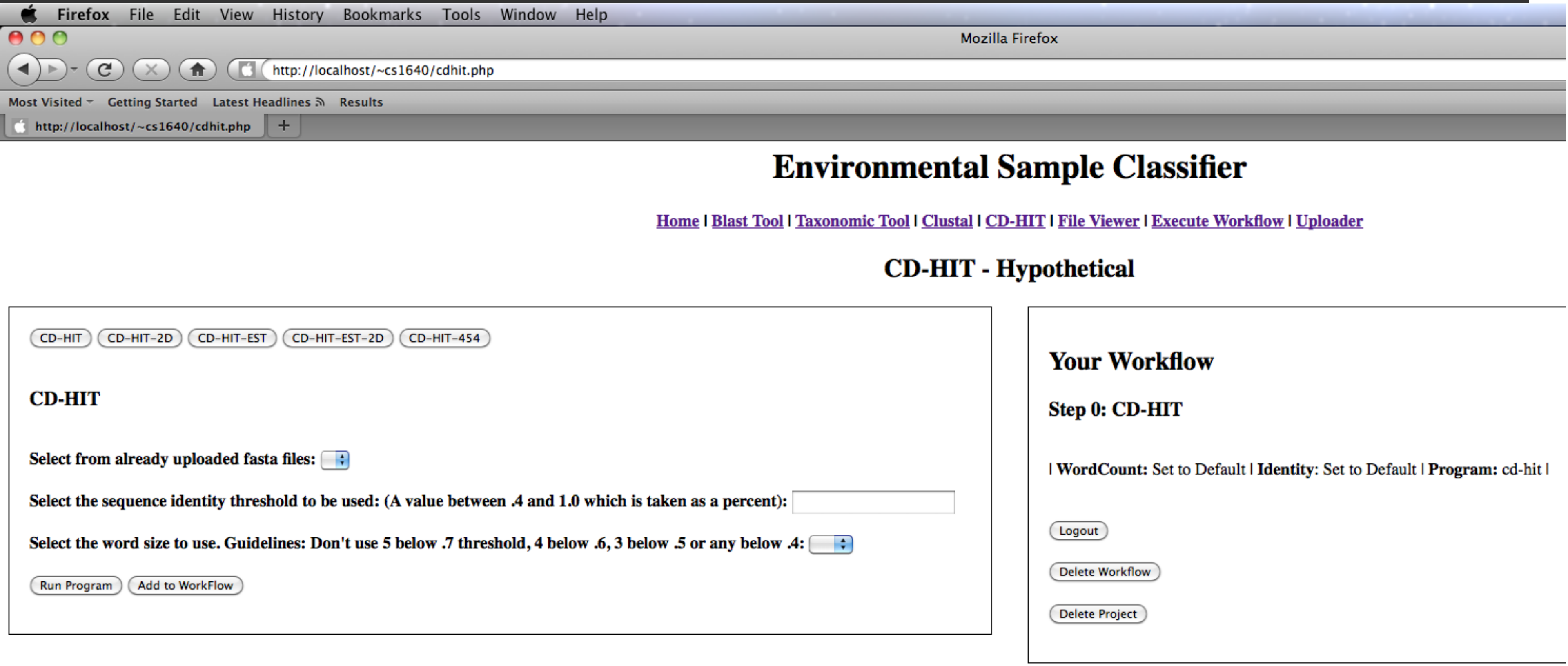
Using the Website

Create a New Project



Select project from the list or start new one Hypothetical Test Submit Query

Add CD-HIT to Workflow



Environmental Sample Classifier

Home | Blast Tool | Taxonomic Tool | Clustal | CD-HIT | File Viewer | Execute Workflow | Uploader

CD-HIT - Hypothetical

CD-HIT

Select from already uploaded fasta files:

Select the sequence identity threshold to be used: (A value between .4 and 1.0 which is taken as a percent):

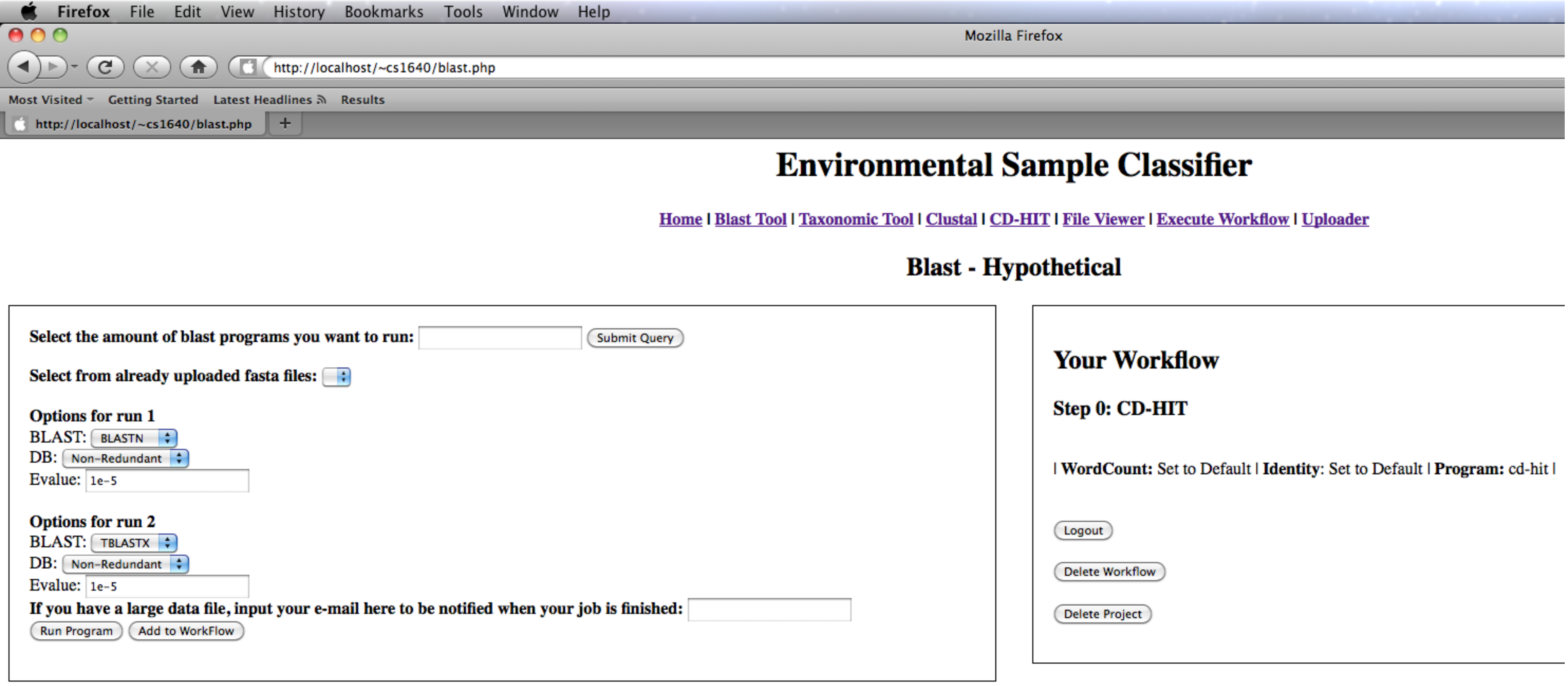
Select the word size to use. Guidelines: Don't use 5 below .7 threshold, 4 below .6, 3 below .5 or any below .4:

Your Workflow

Step 0: CD-HIT

| **WordCount:** Set to Default | **Identity**: Set to Default | **Program:** cd-hit |

Add BLAST to the Workflow



Firefox File Edit View History Bookmarks Tools Window Help
Mozilla Firefox
http://localhost/~cs1640/blast.php
Most Visited Getting Started Latest Headlines Results
http://localhost/~cs1640/blast.php

Environmental Sample Classifier

Home | Blast Tool | Taxonomic Tool | Clustal | CD-HIT | File Viewer | Execute Workflow | Uploader

Blast - Hypothetical

Select the amount of blast programs you want to run: Submit Query

Select from already uploaded fasta files:

Options for run 1
BLAST: BLASTN
DB: Non-Redundant
Evalue: 1e-5

Options for run 2
BLAST: TBLASTX
DB: Non-Redundant
Evalue: 1e-5

If you have a large data file, input your e-mail here to be notified when your job is finished:

Run Program   Add to WorkFlow

Your Workflow

Step 0: CD-HIT

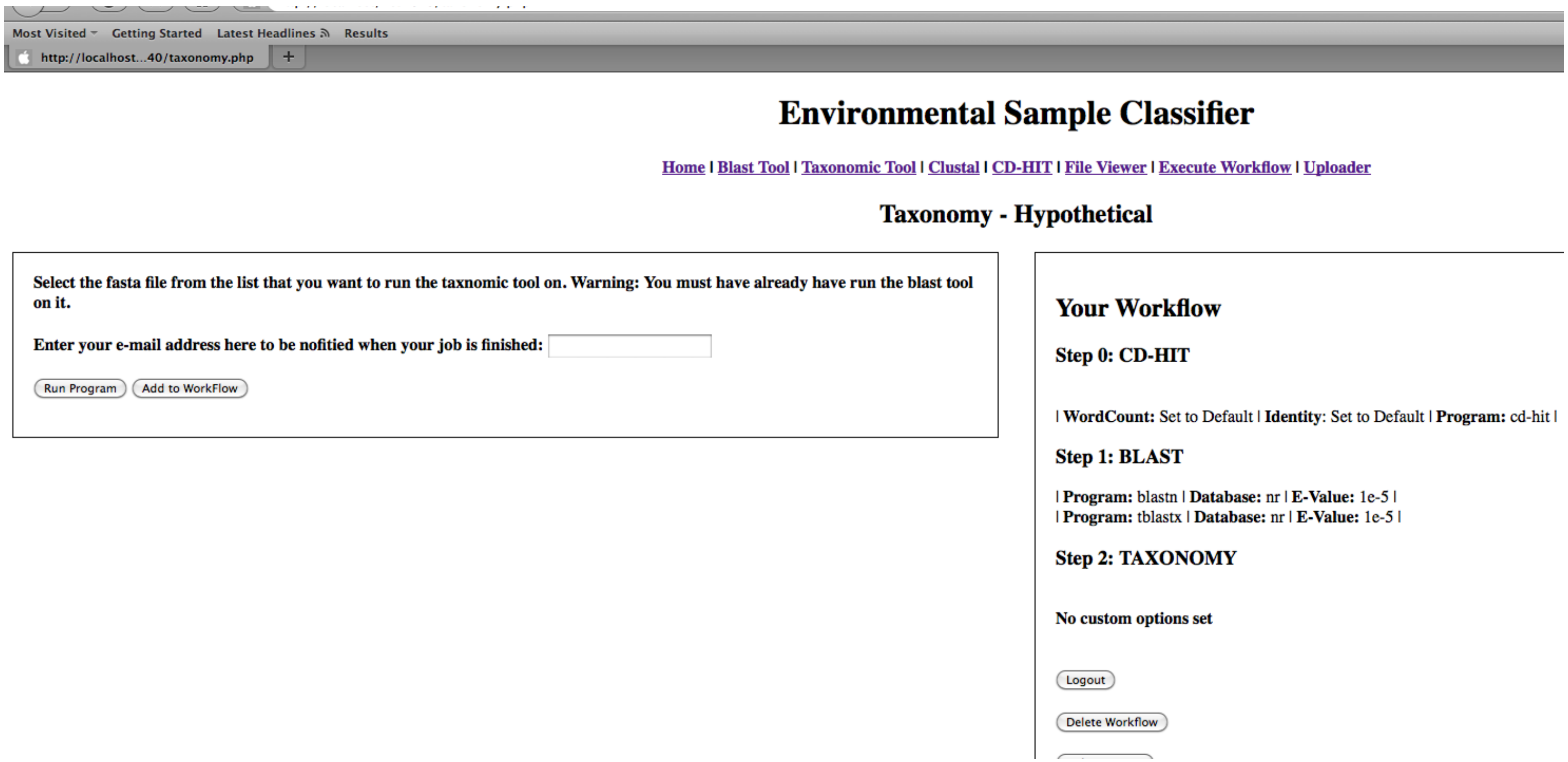
| **WordCount:** Set to Default | **Identity**: Set to Default | **Program:** cd-hit |

Logout

Delete Workflow

Delete Project

Add Taxonomy to Workflow



Environmental Sample Classifier

Home | Blast Tool | Taxonomic Tool | Clustal | CD-HIT | File Viewer | Execute Workflow | Uploader

Taxonomy - Hypothetical

Select the fasta file from the list that you want to run the taxnomic tool on. Warning: You must have already have run the blast tool on it.

Enter your e-mail address here to be nofitied when your job is finished:

Your Workflow

Step 0: CD-HIT

| **WordCount:** Set to Default | **Identity**: Set to Default | **Program:** cd-hit |

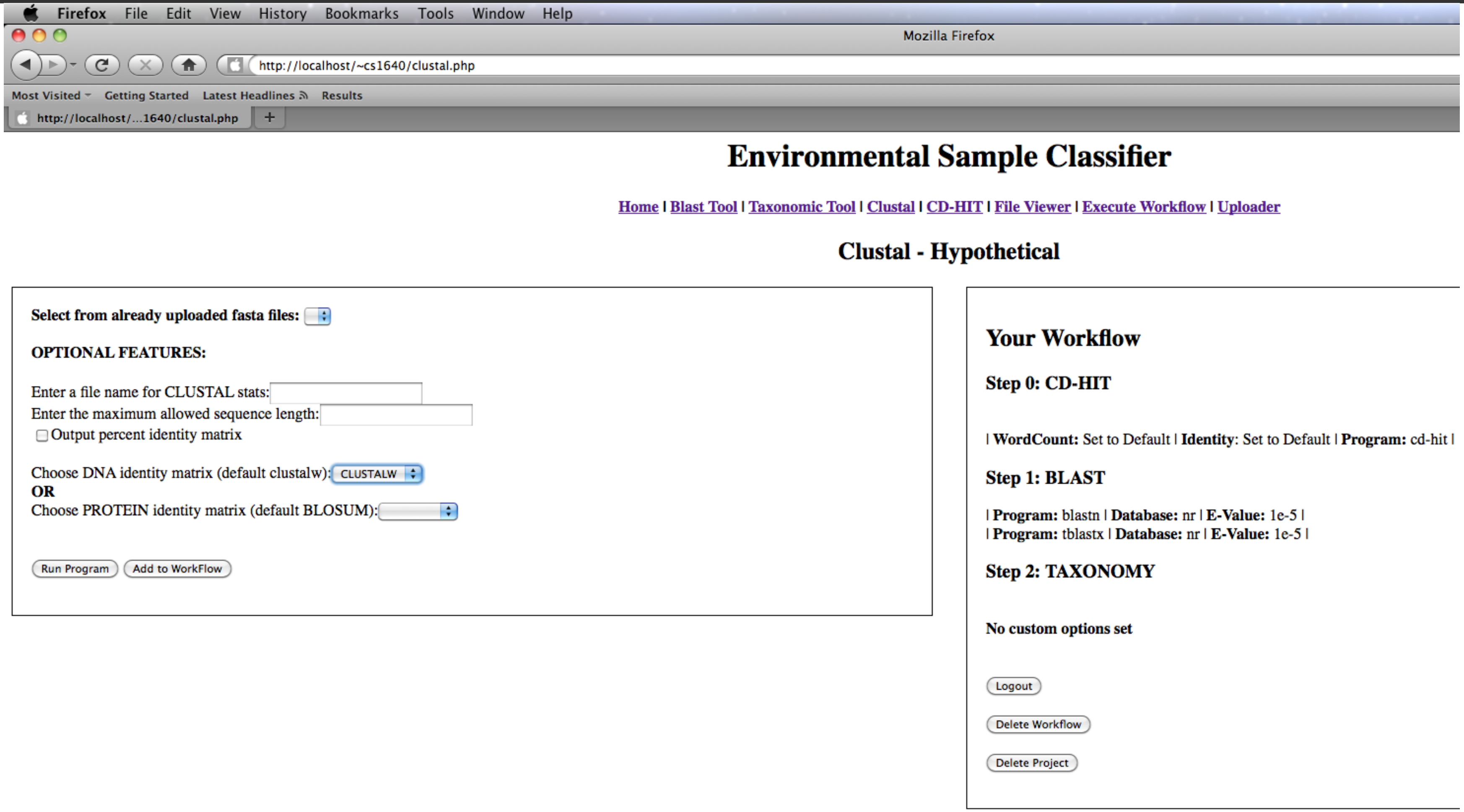
Step 1: BLAST

| **Program:** blastn | **Database:** nr | **E-Value:** 1e-5 |
| **Program:** tblastx | **Database:** nr | **E-Value:** 1e-5 |

Step 2: TAXONOMY

No custom options set

Add Clustal to Workflow


Firefox File Edit View History Bookmarks Tools Window Help
Mozilla Firefox
http://localhost/~cs1640/clustal.php
Most Visited Getting Started Latest Headlines Results
http://localhost/...1640/clustal.php

Environmental Sample Classifier

Home | Blast Tool | Taxonomic Tool | Clustal | CD-HIT | File Viewer | Execute Workflow | Uploader

Clustal - Hypothetical

Select from already uploaded fasta files:

OPTIONAL FEATURES:

Enter a file name for CLUSTAL stats:
Enter the maximum allowed sequence length:
□ Output percent identity matrix

Choose DNA identity matrix (default clustalw): CLUSTALW
OR
Choose PROTEIN identity matrix (default BLOSUM):

Your Workflow

Step 0: CD-HIT

| **WordCount:** Set to Default | **Identity**: Set to Default | **Program:** cd-hit |

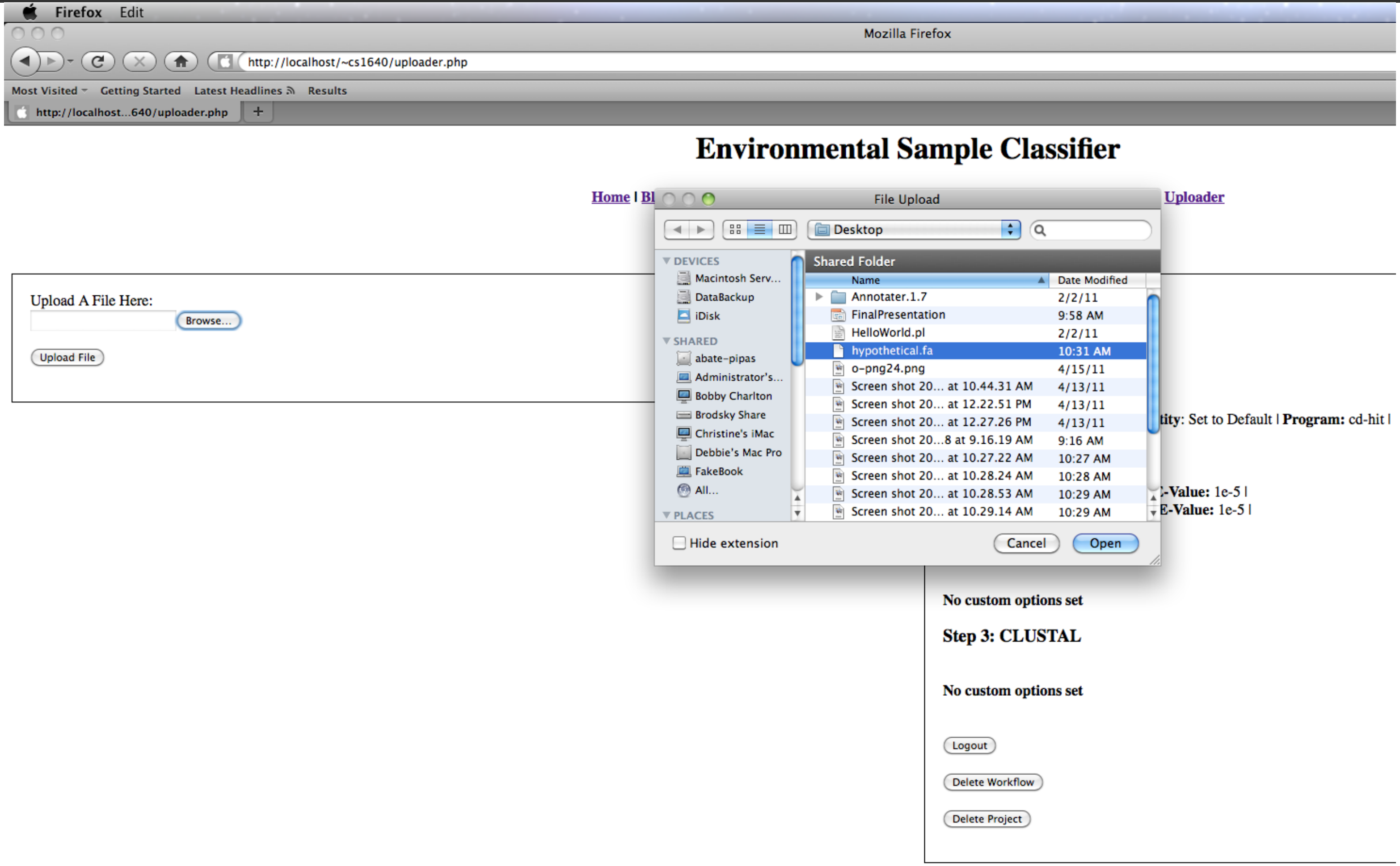
Step 1: BLAST

| **Program:** blastn | **Database:** nr | **E-Value:** 1e-5 |
| **Program:** tblastx | **Database:** nr | **E-Value:** 1e-5 |

Step 2: TAXONOMY

No custom options set

Upload our Fasta File



http://localhost/~cs1640/uploader.php

Environmental Sample Classifier

Upload A File Here:

Name	Date Modified
Annotater.1.7	2/2/11
FinalPresentation	9:58 AM
HelloWorld.pl	2/2/11
hypothetical.fa	10:31 AM
o-png24.png	4/15/11
Screen shot 20... at 10.44.31 AM	4/13/11
Screen shot 20... at 12.22.51 PM	4/13/11
Screen shot 20... at 12.27.26 PM	4/13/11
Screen shot 20...8 at 9.16.19 AM	9:16 AM
Screen shot 20... at 10.27.22 AM	10:27 AM
Screen shot 20... at 10.28.24 AM	10:28 AM
Screen shot 20... at 10.28.53 AM	10:29 AM
Screen shot 20... at 10.29.14 AM	10:29 AM

□ Hide extension

tity: Set to Default | **Program:** cd-hit |

-Value: 1e-5 |
E-Value: 1e-5 |

No custom options set

Step 3: CLUSTAL

No custom options set

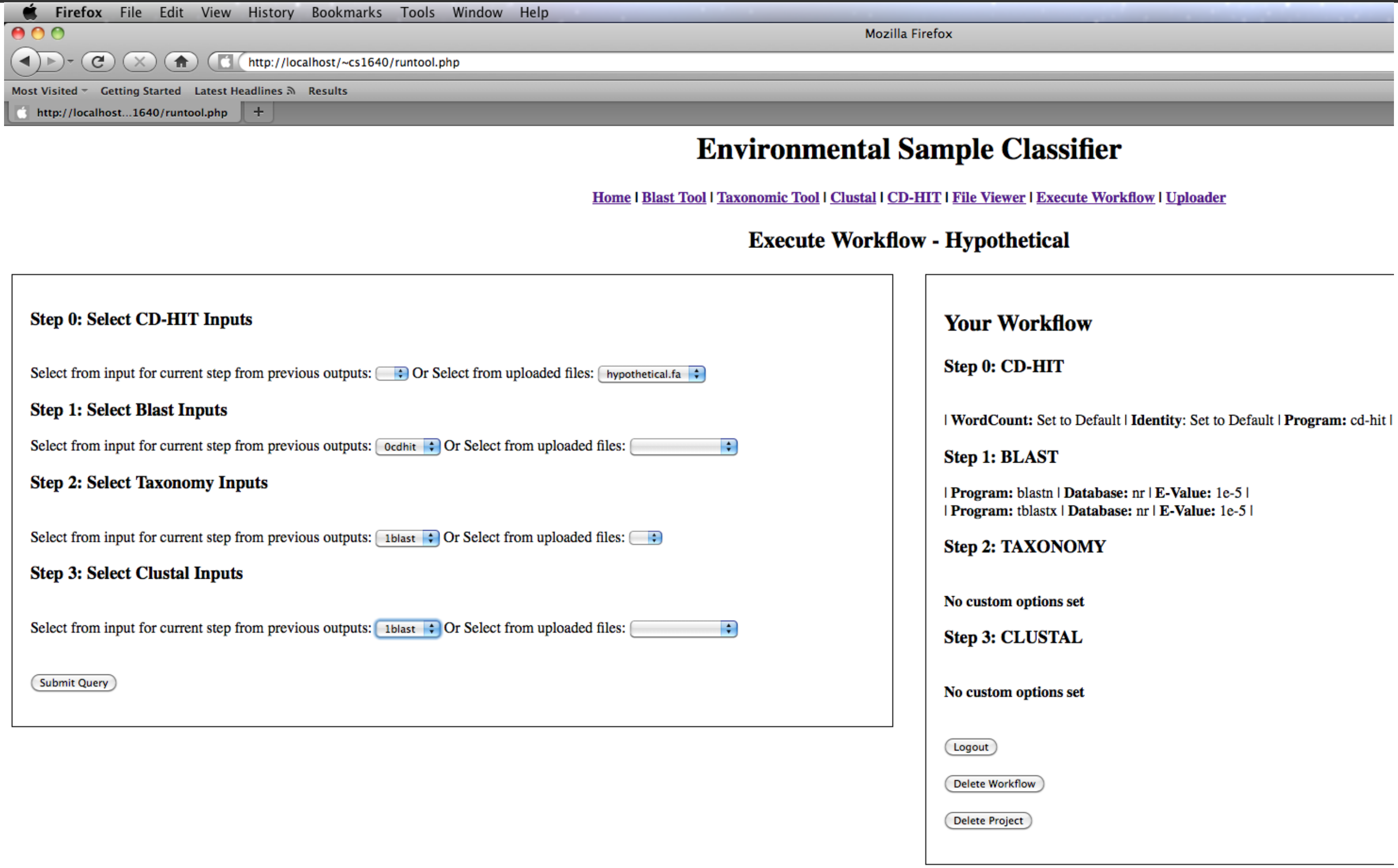
Example Fasta File

- Fasta refers to the specific text format for biological sequence data.
- >Hypothetical.ID Comment

```
ATCGATCGATCGATCGATCGATCGATCGATACGCTAGACTA
CGACTACGACTACGATCACGACTACGACTACGACTACGA
CTACGACTACGACTACGACTACGATCAGCTATACGATCAG
CTACGATCAGCTAGACTAGACTACGACTACGATCGATCAG
CATCAGCATCGATCGATCGATCGACTACGCA
```

etc...

Run the Workflow


Firefox File Edit View History Bookmarks Tools Window Help
Mozilla Firefox
http://localhost/~cs1640/runtool.php
Most Visited Getting Started Latest Headlines Results
http://localhost...1640/runtool.php

Environmental Sample Classifier

Home | Blast Tool | Taxonomic Tool | Clustal | CD-HIT | File Viewer | Execute Workflow | Uploader

Execute Workflow - Hypothetical

Step 0: Select CD-HIT Inputs

Select from input for current step from previous outputs: [ ] Or Select from uploaded files: [hypothetical.fa]

Step 1: Select Blast Inputs

Select from input for current step from previous outputs: [0cdhit] Or Select from uploaded files: [ ]

Step 2: Select Taxonomy Inputs

Select from input for current step from previous outputs: [1blast] Or Select from uploaded files: [ ]

Step 3: Select Clustal Inputs

Select from input for current step from previous outputs: [1blast] Or Select from uploaded files: [ ]

Your Workflow

Step 0: CD-HIT

| **WordCount:** Set to Default | **Identity**: Set to Default | **Program:** cd-hit |

Step 1: BLAST

| **Program:** blastn | **Database:** nr | **E-Value:** 1e-5 |
| **Program:** tblastx | **Database:** nr | **E-Value:** 1e-5 |

Step 2: TAXONOMY

No custom options set

Step 3: CLUSTAL

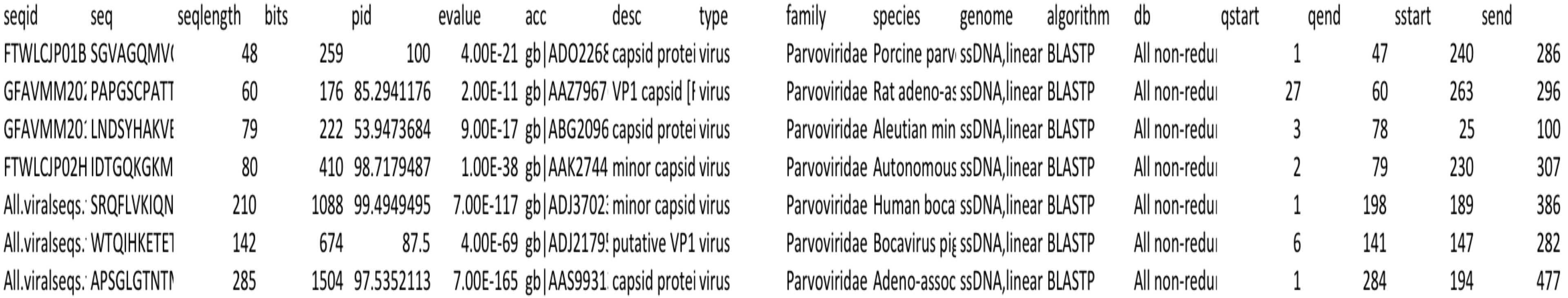
No custom options set

Example Blast Output

```
TBLASTX 2.2.24+


Reference: Stephen F. Altschul, Thomas L. Madden,
Alejandro A.
Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and
David J.
Lipman (1997), "Gapped BLAST and PSI-BLAST: a new
generation of
protein database search programs", Nucleic Acids Res.
25:3389-3402.


Database: All GenBank+EMBL+DDBJ+PDB sequences (but no EST,
STS,
GSS,environmental samples or phase 0, 1 or 2 HTGS
sequences)
           13,837,274 sequences; 33,786,494,993 total
letters


Query=  GFAVMM201ENU6Z
Length=152


RID: 1288023192-8524-201920965474.BLASTQ4

Score     E
Sequences producing significant alignments:
(Bits)  Value  N


gb|L32166.1|BYTV1  Banana bunchy top virus (BBTV DNA I) V1
and C1...  63.8    1e-08  1
gb|GQ404856.1|  Human stool-associated circular virus
NG13, compl...  62.9    2e-08  1
gb|EU430730.1|  Banana bunchy top virus putative satellite
4, com...  62.5    3e-08  1
gb|AF416471.1|  Banana bunchy top virus putative satellite
3 DNA ...  62.5    3e-08  1
gb|AF216222.1|AF216222  Banana bunchy top virus satellite
S2 repl...  61.6    5e-08  1



>gb|L32166.1|BYTV1 Banana bunchy top virus (BBTV DNA I) V1
and C1-C3 genes, complete
cds's
Length=1106

 Score = 63.8 bits (133),  Expect = 1e-08
 Identities = 24/47 (51%), Positives = 32/47 (68%), Gaps =
0/47 (0%)
 Frame = +3/+2
```

```
Query  3    GTRHYQGFLILKKRNRMTWLKSNINNRAHWEKTRGTDKQAADYCRKD  143
            G +H QG+L LKKR R+  LK    +RAHWE  RGTD++ + YC K+
Sbjct  197  GQKHLQGYLSLKKRIRLGGLKKKYGSRAHWEIARGTDEENSKYCSKE  337


>gb|GQ404856.1| Human stool-associated circular virus NG13,
complete genome
Length=1699

 Score = 62.9 bits (131),  Expect = 2e-08
 Identities = 24/46 (52%), Positives = 34/46 (73%), Gaps = 0/46
(0%)
 Frame = +3/+1

Query  3    GTRHYQGFLILKKRNRMTWLKSNINNRAHWEKTRGTDKQAADYCRK  140
            GT H QGF  LKK+ R+T LK+ +N+RAH+E+ +G+D+Q   YC K
Sbjct  181  GTPHLQGFFNLKKKKRLTSLKAWLNDRAHYEEAKGSDEQNRRYCSK  318


>gb|EU430730.1| Banana bunchy top virus putative satellite 4,
complete sequence
Length=1103

 Score = 62.5 bits (130),  Expect = 3e-08
 Identities = 23/47 (48%), Positives = 32/47 (68%), Gaps = 0/47
(0%)
 Frame = +3/+2

Query  3    GTRHYQGFLILKKRNRMTWLKSNINNRAHWEKTRGTDKQAADYCRKD  143
            G +H QG+L LKKR R++ +K   ++RAHWEK RG+D     YC K+
Sbjct  179  GRKHLQGYLSLKKRFRISGIKKKYSSRAHWEKARGSDYDNKAYCSKE  319


>gb|AF416471.1| Banana bunchy top virus putative satellite 3 DNA
molecule, complete
sequence
Length=1100

 Score = 62.5 bits (130),  Expect = 3e-08
 Identities = 23/47 (48%), Positives = 32/47 (68%), Gaps = 0/47
(0%)
 Frame = +3/+2

Query  3    GTRHYQGFLILKKRNRMTWLKSNINNRAHWEKTRGTDKQAADYCRKD  143
            G +H QG+L LKKR R++ +K   ++RAHWEK RG+D     YC K+
Sbjct  179  GRKHLQGYLSLKKRFRISGIKKKYSSRAHWEKARGSDYDNKAYCSKE  319
```

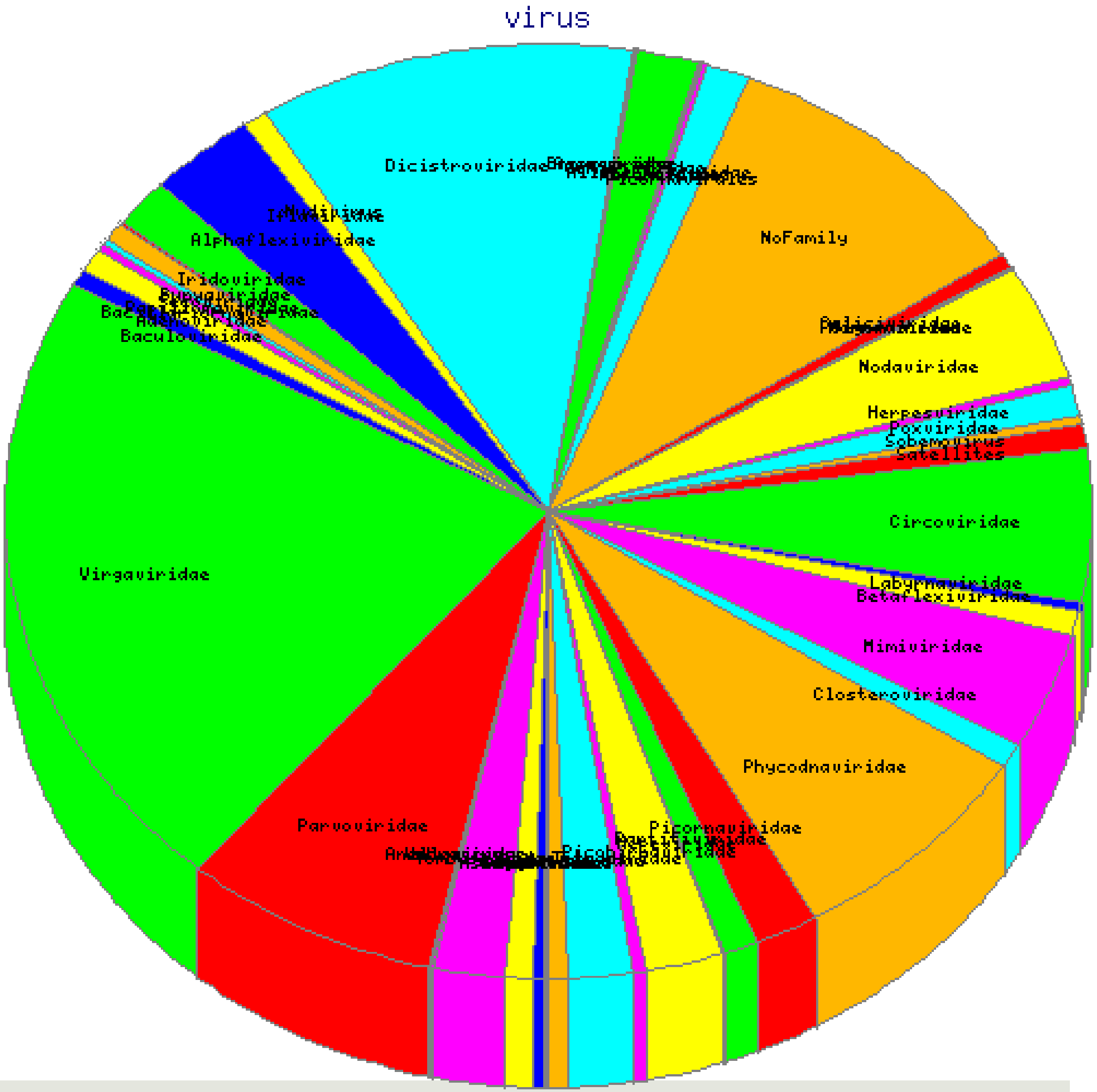
Example Taxonomic Report



seqid	seq	seqlength	bits	pid	evalue	acc	desc	type	family	species	genome	algorithm	db	qstart	qend	sstart	send
FTWLCJP01B	SGVAGQMV(	48	259	100	4.00E-21	gb|ADO2268	capsid protei	virus	Parvoviridae	Porcine parv	ssDNA,linear	BLASTP	All non-redu	1	47	240	286
GFAVMM202	PAPGSCPATT	60	176	85.2941176	2.00E-11	gb|AAZ7967	VP1 capsid [F	virus	Parvoviridae	Rat adeno-as	ssDNA,linear	BLASTP	All non-redu	27	60	263	296
GFAVMM201	LNDSYHAKVE	79	222	53.9473684	9.00E-17	gb|ABG2096	capsid protei	virus	Parvoviridae	Aleutian min	ssDNA,linear	BLASTP	All non-redu	3	78	25	100
FTWLCJP02H	IDTGQKGKM	80	410	98.7179487	1.00E-38	gb|AAK2744	minor capsid	virus	Parvoviridae	Autonomous	ssDNA,linear	BLASTP	All non-redu	2	79	230	307
All.viralseqs.	SRQFLVKIQN	210	1088	99.4949495	7.00E-117	gb|ADJ3702	minor capsid	virus	Parvoviridae	Human boca	ssDNA,linear	BLASTP	All non-redu	1	198	189	386
All.viralseqs.	WTQIHKETET	142	674	87.5	4.00E-69	gb|ADJ2179	putative VP1	virus	Parvoviridae	Bocavirus pig	ssDNA,linear	BLASTP	All non-redu	6	141	147	282
All.viralseqs.	APSGLGTNTN	285	1504	97.5352113	7.00E-165	gb|AAS9931	capsid protei	virus	Parvoviridae	Adeno-assoc	ssDNA,linear	BLASTP	All non-redu	1	284	194	477

Example Taxonomic Output

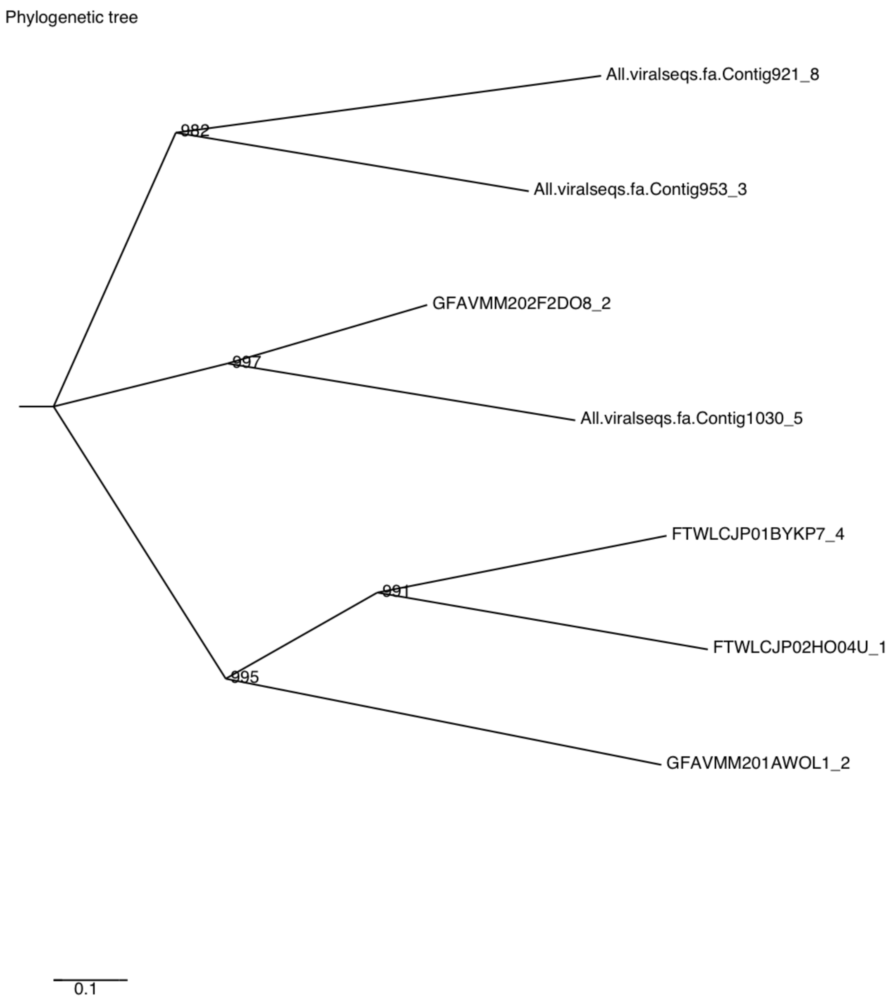
Potyviridae	2
Luteoviridae	1
Ourmiavirus	7
Astroviridae	21
Tombusviridae	50
Umbravirus	2
Anelloviridae	1
Parvoviridae	182
Virgaviridae	519
Adenoviridae	19
Baculoviridae	15
Papillomaviridae	5
Bacillariornaviridae	6
Secoviridae	15
Alphaflexiviridae	76
Bunyaviridae	2
Iridoviridae	41
Iflaviridae	17
Dicistroviridae	266
Nudivirus	1
Bromoviridae	2
Ascoviridae	1
Alloherpesviridae	2



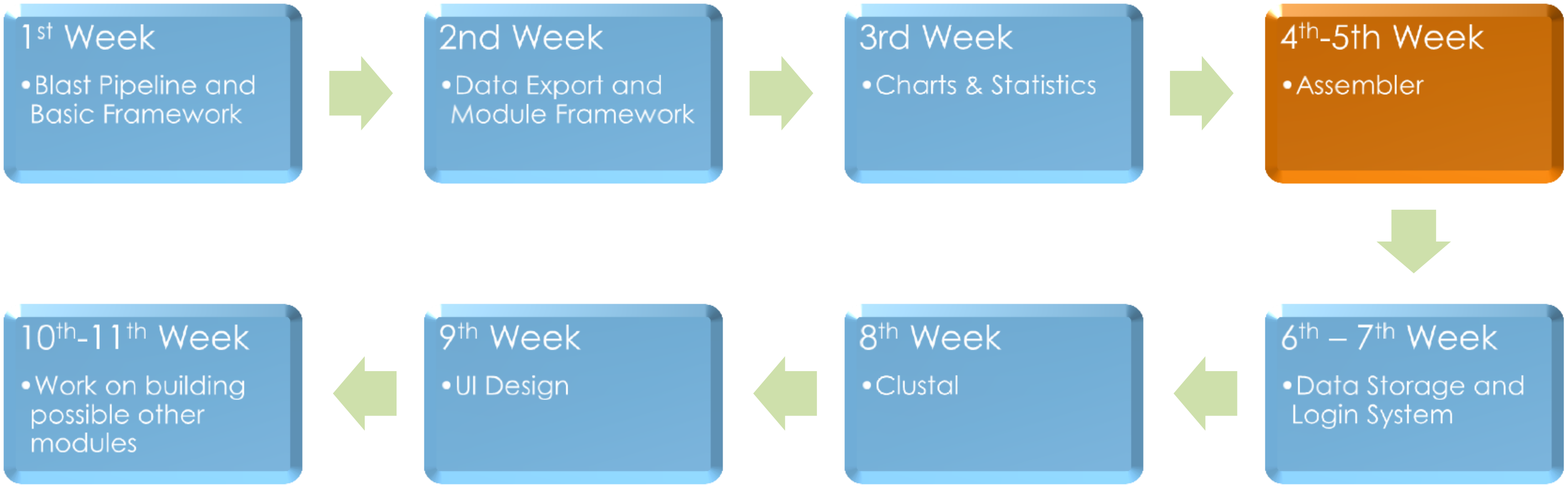
Example Clustal Alignment

```
CLUSTAL 2.1 multiple sequence alignment

All.viralseqs.fa.Contig921_8       QDWQRLTNEYKRFRPKGMHVKIYNLQIKQILSNGADVTYNNDLTAGVHIF 100

All.viralseqs.fa.Contig953_3       NDWQRLLNNYKKWRPQKMRVQLYNLQIKQVVKLGTDTLYNNDLTAGVHVM 87

GFAVMM202F2DO8_2                   RDWQRLVN------------------------------------------ 60

All.viralseqs.fa.Contig1030_5      RDWQRLINNNWGFRPKRLNFKLFNIQVKEVTQNDGTTTIANNLTSTVQVF 150

FTWLCJP01BYKP7_4                   ADWQLISNNMTEIT------------------------------------ 48

FTWLCJP02HO04U_1                   SDWQFIQNSMESLNPESFSQELFNVVVKMVTEQDIAGTTTKVYK------ 80

GFAVMM201AWOL1_2                   ADWQQTITTCRNLEPIHLHQSIDNIVIKTVTKQGTGAEETTQYNNDLTAH 77

                                    ***   .
```

Example Clustal Phylogenetic Tree



Original Goals for the Semester



1st Week
- Blast Pipeline and Basic Framework

2nd Week
- Data Export and Module Framework

3rd Week
- Charts & Statistics

4th-5th Week
- Assembler

6th – 7th Week
- Data Storage and Login System

8th Week
- Clustal

9th Week
- UI Design

10th-11th Week
- Work on building possible other modules

Conclusion

- Molecular Biologists need comprehensive tools for analyzing metagenomic samples.
- While the tools exist each of them are not comprehensive and we assembled the ones we think are useful.
- Our collection of assembled tools is designed to analyze the statistical similarity of sequences in a variety of methods.
- These tools will help in the identification of novel viral sequences.