Tutorial: Motif finders

BIOSC 1540 Fall 2018

0. Using Secure FTP to transfer files between the CRC cluster and your computer

You will need a Secure FTP (SFTP) program to work with files off of the CRC.

For Mac OS X, Fetch should be available through the Pitt Software Distribution Services. Alternatives include FileZilla https://filezilla-project.org/ and CyberDuck https://cyberduck.en.softonic.com/mac

For Windows, WinSCP is available through the Pitt Software Distribution Services or at https://winscp.net/eng/index.php

If you need assistance downloading / installing / using SFTP software, please ask your local Computing Support staff member or visit the Walk-In Support Desk at the University Store on Fifth.

When you run the SFTP program, it will ask you for the host name (`h2p.crc.pitt.edu`) as well as your Pitt username/password. Make sure you are connecting using the SFTP protocol, not FTP (which is not secure). You will also need to be running Pulse Secure.

SFTP programs all have an interface that shows you the directory tree of the remote computer (in this case, the CRC). By default, the program should show you the contents of your home folder on the CRC. You can navigate to a folder of interest by double clicking successive subdirectories. You can transfer a file or a folder by clicking on it and selecting the "Get" function either as a menu option or menu icon; or, with most programs you should be able to click and drag the file/folder onto your desktop.

I. Using homer

1. Open an interactive CRC session (here, 2 hours) and load the course module

   ```
   $ crc-interactive.py -s -n 1 -c 1 -t 2
   $ module load teaching/biosc1540-2018f
   ```

2. Generate a FASTA file of interest, e.g. ChIP-seq peaks, using `bedtools getfasta` and other tools as necessary. Follow the advice from lecture to make your FASTA file in a way to maximize motif finding sensitivity.
3. Run the homer motif finding command

   ```
   $ findMotifs.pl my_peaks.fa fasta output_dir -mset vert
   ```

   Your FASTA file is called `my_peaks.fa` and you are searching for vertebrate motifs (`-mset vert`). The result files will be saved into a directory called `output_dir`. This is the default way to run homer, which will perform both a database scan as well as de novo motif discovery. As such, it will potentially take a long time to run (20-30 minutes) depending on the size of your input file. You can disable de novo motif discovery with the `-nomotif` flag

```
$ findMotifs.pl my_peaks.fa fasta output_dir -mset vert -nomotif
```

4. Once the program is done, the output files will be in `output_dir` or whatever directory name you specified when you ran the command. Use your SFTP program to log into the cluster (`h2p.crc.pitt.edu`) and transfer the entire folder to your computer.
5. To look at the results of the motif scanning, open the `knownResults.html` file, which is just a local web page (double clicking should open the file in your default Web browser). Motifs are listed in descending order according to significance (p-value). If you see a red asterisk (*) that means you should not consider that motif to be a true positive (i.e., the corrected significance q-value did not pass the threshold).
6. To look at the results of the de novo motif discovery, assuming you ran it, open the `homerResults.html` file. The column named "Best Match/Details" comes from comparing the motif to a database (not the same as motif scanning) and looking for the best match. The name of the transcription factor listed there may or may not be meaningful in your context, but it does implicate a likely transcription factor family that recognizes your motif. Feel free to click on the "More Information" link to view additional database hits.
7. The motifs are rendered in an image format called SVG, which may be difficult to work with. You can always take a screenshot to use as part of your figure.

II. Using MEME-ChIP

1. Generate a FASTA file of interest, e.g. ChIP-seq peaks, using `bedtools getfasta` and other tools as necessary
2. Use your SFTP program to transfer that FASTA file to your local computer
3. Navigate to http://meme-suite.org/tools/meme-chip
4. Under "Input the primary sequences" upload your FASTA file using the Browse button. If the FASTA file is too big, you will get a warning message (you may want to limit your FASTA file to no more than 500-1000 sequences anyway) (See Fig 1)
5. Click the button to "Start Search"
6. Wait. The web page will automatically reload as the job is running, which can potentially be a long time (hours) due to server load. You can bookmark the web page and come back to it later. You can also try to use the Alternate Server linked on the left panel.
7. When MEME-ChIP is done, the web page will look like Fig 2. Click on the "MEME-ChIP HTML output" link for the main results page.
8. Individual motifs are image files that you can right-click to save to your computer. If you do this for your figures, make sure to record the E-value as well, which is similar to a p-value (lower means better)

Figure 1

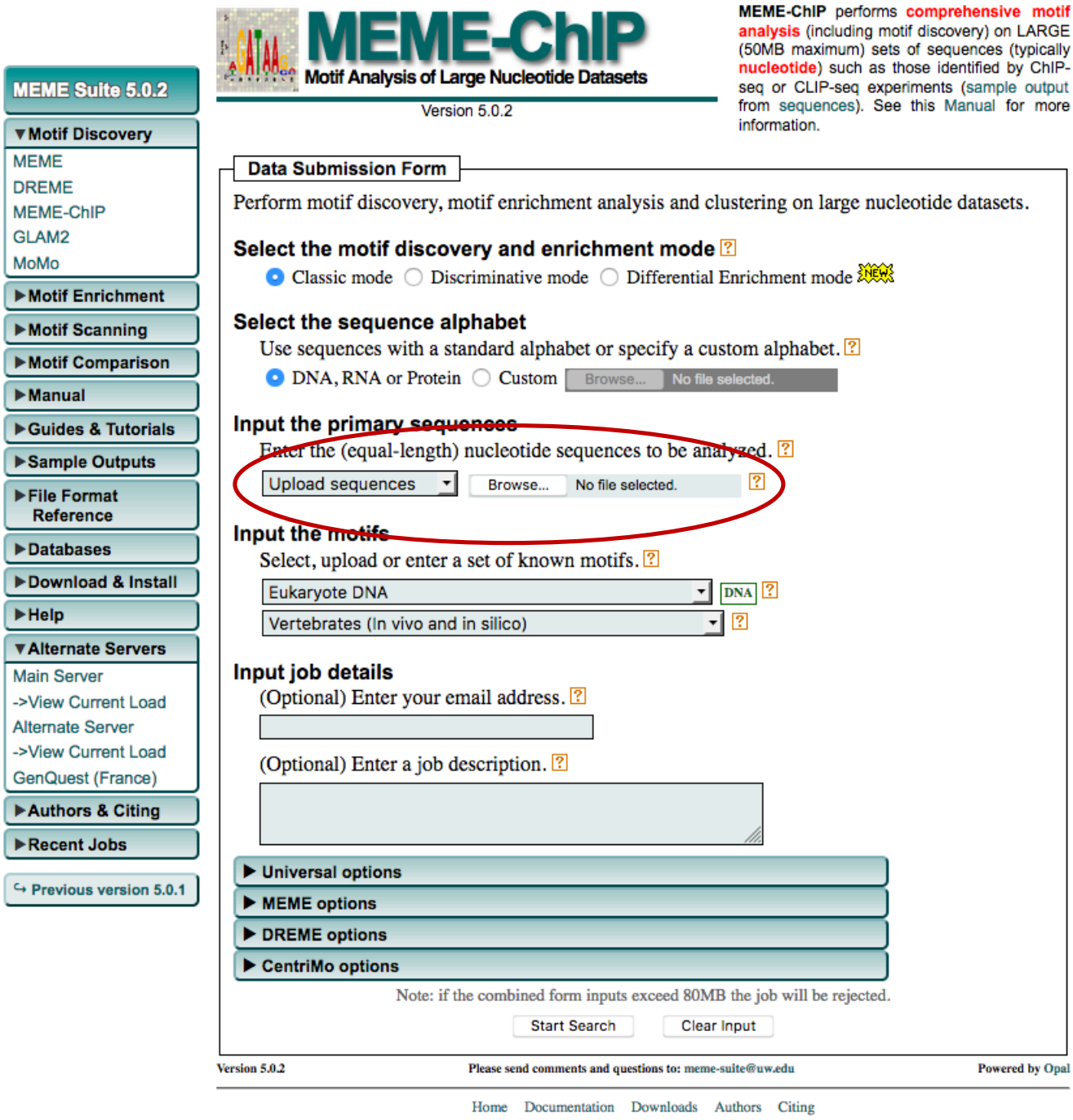
MEME Suite 5.0.2
▼Motif Discovery
MEME
DREME
MEME-ChIP
GLAM2
MoMo
▶Motif Enrichment
▶Motif Scanning
▶Motif Comparison
▶Manual
▶Guides & Tutorials
▶Sample Outputs
▶File Format Reference
▶Databases
▶Download & Install
▶Help
▼Alternate Servers
Main Server
->View Current Load
Alternate Server
->View Current Load
GenQuest (France)
▶Authors & Citing
▶Recent Jobs
↪ Previous version 5.0.1
MEME-ChIP
Motif Analysis of Large Nucleotide Datasets
Version 5.0.2
MEME-ChIP performs comprehensive motif analysis (including motif discovery) on LARGE (50MB maximum) sets of sequences (typically nucleotide) such as those identified by ChIP-seq or CLIP-seq experiments (sample output from sequences). See this Manual for more information.
Data Submission Form
Perform motif discovery, motif enrichment analysis and clustering on large nucleotide datasets.
Select the motif discovery and enrichment mode
Classic mode
Discriminative mode
Differential Enrichment mode
NEW
Select the sequence alphabet
Use sequences with a standard alphabet or specify a custom alphabet.
DNA, RNA or Protein
Custom
Browse...
No file selected.
Input the primary sequences
Enter the (equal-length) nucleotide sequences to be analyzed.
Upload sequences
Browse...
No file selected.
Input the motifs
Select, upload or enter a set of known motifs.
Eukaryote DNA
DNA
Vertebrates (In vivo and in silico)
Input job details
(Optional) Enter your email address.
(Optional) Enter a job description.
▶ Universal options
▶ MEME options
▶ DREME options
▶ CentriMo options
Note: if the combined form inputs exceed 80MB the job will be rejected.
Start Search
Clear Input
Version 5.0.2
Please send comments and questions to: meme-suite@uw.edu
Powered by Opal
Home
Documentation
Downloads
Authors
Citing

Figure 2


MEME-ChIP
Motif Analysis of Large Nucleotide Datasets
MEME Suite 5.0.2
▼Motif Discovery
MEME
DREME
MEME-ChIP
GLAM2
MoMo
▶Motif Enrichment
▶Motif Scanning
▶Motif Comparison
▶Manual
▶Guides & Tutorials
▶Sample Outputs
▶File Format Reference
▶Databases
▶Download & Install
▶Help
▼Alternate Servers
Main Server
->View Current Load
Alternate Server
->View Current Load
GenQuest (France)
▶Authors & Citing
▼Recent Jobs
MEME-ChIP 11:43 AM
Clear All
Previous version 5.0.1
Your MEME-ChIP job is complete. The results should be displayed below.
Job Details ...
Results
MEME-ChIP HTML output
Gzipped Tar of all output
MEME-ChIP TSV output
MEME-ChIP motif output
Uploaded Sequences
Messages
Status Messages
Arguments ok
Starting MEME-ChIP
meme-chip -oc . -time 300 -ccut 100 -order 1 -db db/EUKARYOTE/jolma2013.meme -db db/JASPAR/JASPAR2018_
minw 6 -meme-maxw 30 -meme-nmotifs 3 -meme-searchsize 100000 -dreme-e 0.05 -centrimo-score 5.0 -centri
MEME-ChIP is starting subprocess getsize
getsize ./pou_peaks_1000.fa 1> $metrics
MEME-ChIP subprocess getsize ran successfully in 0.0 seconds
MEME-ChIP is starting subprocess fasta-most
fasta-most -min 50 < ./pou_peaks_1000.fa 1> $metrics
MEME-ChIP subprocess fasta-most ran successfully in 0.1 seconds
MEME-ChIP is starting subprocess fasta-center
fasta-center -dna -len 100 < ./pou_peaks_1000.fa 1> ./seqs-centered
MEME-ChIP subprocess fasta-center ran successfully in 0.2 seconds
MEME-ChIP is starting subprocess fasta-shuffle-letters
fasta-shuffle-letters ./seqs-centered ./seqs-shuffled -kmer 2 -tag -dinuc -dna -seed 1
MEME-ChIP subprocess fasta-shuffle-letters ran successfully in 0.0 seconds
MEME-ChIP is starting subprocess fasta-get-markov
fasta-get-markov -nostatus -nosummary -dna -m 1 ./pou_peaks_1000.fa ./background
MEME-ChIP subprocess fasta-get-markov ran successfully in 0.0 seconds
MEME-ChIP is starting subprocess meme
meme ./seqs-centered -oc meme_out -mod zoops -nmotifs 3 -minw 6 -maxw 30 -bfile ./background -dna -ses
MEME-ChIP subprocess meme ran successfully in 2581.9 seconds
MEME-ChIP is starting subprocess dreme
dreme -verbosity 1 -oc dreme_out -png -dna -p ./seqs-centered -n ./seqs-shuffled -t 6106 -e 0.05
MEME-ChIP subprocess dreme ran successfully in 196.2 seconds
MEME-ChIP is starting subprocess centrimo
centrimo -seqlen 252 -verbosity 1 -oc centrimo_out -bfile ./background -score 5.0 -ethresh 10.0 ./pou_