CENTER FOR
GENOME RESEARCH &
BIOCOMPUTING

"Introduction to Unix/Linux"
INX_U18, Day 6, 2018-08-06

Installing binaries, uname, hmmer and muscle, public data (wget and sftp)

Learning Outcome(s):

Install and run software from your home directory.
Download bioinformatics data sets and analysis programs from internet sources, and decompress them if necessary.

Matthew Peterson, OSU CGRB, matthew@cgrb.oregonstate.edu

Installation of HMMER

- Last class we downloaded the **source** code of HMMER (hmmer.org) and performed a "canonical" install into our `$HOME/local/bin`

```
./configure          # Configure (with --prefix)
make                 # Build
make check           # Automated tests
make install         # Automated install
```

Oregon State UNIVERSITY

What does HMMER do?

- Given a "**m**ultiple **s**equence **a**lignment" (**MSA**) of protein sequences, e.g., the same gene in multiple species, it build a **H**idden **M**arkov **M**odel, i.e. an **HMM** "profile" of them.
- This profile serves as a probabilistic model of the whole gene set, so that later that profile can be compared to other sequences to find matches ("in the database").

https://en.wikipedia.org/wiki/**M**ultiple_**s**equence_**a**lignment

Oregon State UNIVERSITY

HMMER workflow

Input: Query Sequence Set

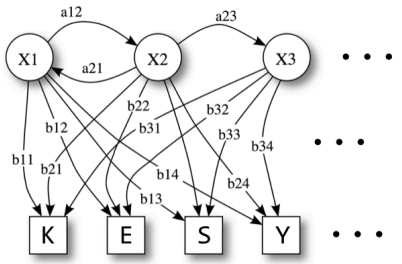
```
...SKEAEYLVKQLNTVME...
...SKEAKYLIQQLDTVMK...
...SKERYAAISMFMK...
...AKEGEYLYSNMLNAVMK...
```

?

Multiple Alignment

```
...SKEAEYLVK-QLNTVME...
...SKEAKYLIQ-QLDTVMK...
...SKERYAA----ISMFMK...
...AKEGEYLYSNMLNAVMK...
```

hmmbuild



HMM Profile

Input: Target Sequence Set

```
...CMSDKPDLSEVETFDKSKLTIQQEKEYNQRS...
...SCALEEHVSKEAEYLVKMLNAVMKVTGSFDP...
...DRSQNPPQSKGCCFVTFYTRKAALEAQNALH...
...KMPKDKERSLNPAAAQRKLDKQKSLKKGKAE...
...
```

hmmsearch

SKEAEYLVKMLNAVMKV

Output: Resulting Match

http://library.open.oregonstate.edu/computationalbiology/chapter/installing-bioinformatics-software/

Recommended activities

- Read the HMMER "**Tutorial**" section that describes turning a multiple-alignment of sequences in to a profile (with **`hmmbuild`**) and searching that profile against the larger set (with **`hmmsearch`**).
- Read the peer-reviewed publication that describes the algorithms implemented by HMMER

Oregon State UNIVERSITY

HMMER limitations

- None of the HMMER programs can produce the multiple alignment from a protein dataset
- We need another program program, **`muscle`**, which we'll get in **binary** not **source** code form.

How would you know any of this?!

1. Read the documentation and help text for tools (and the author's papers describing their tools)
2. Ask your colleagues
3. Search the Internet
4. Read the methods sections of papers
5. Just **start!** When stuck, follow 1 through 4 above. Don't get discouraged, it gets easier with practice…

Oregon State UNIVERSITY

Binary versions of `muscle`

http://www.drive5.com/muscle/downloads.htm

Operating system	Processor	Bits	Executable file
Linux	Intel i86	32	muscle3.8.31_i86linux32.tar.gz
Linux	Intel i86	64	muscle3.8.31_i86linux64.tar.gz
Mac OSX	Intel i86	32	muscle3.8.31_i86darwin32.tar.gz
Mac OSX	Intel i86	64	muscle3.8.31_i86darwin64.tar.gz
Mac	PPC	64	muscle3.8.31_macppc.tar.gz
Windows	Intel i86	32 / 64	muscle3.8.31_i86win32.exe
Windows/Cygwin	Intel i86	32 / 64	muscle3.8.31_i86cygwin32.exe

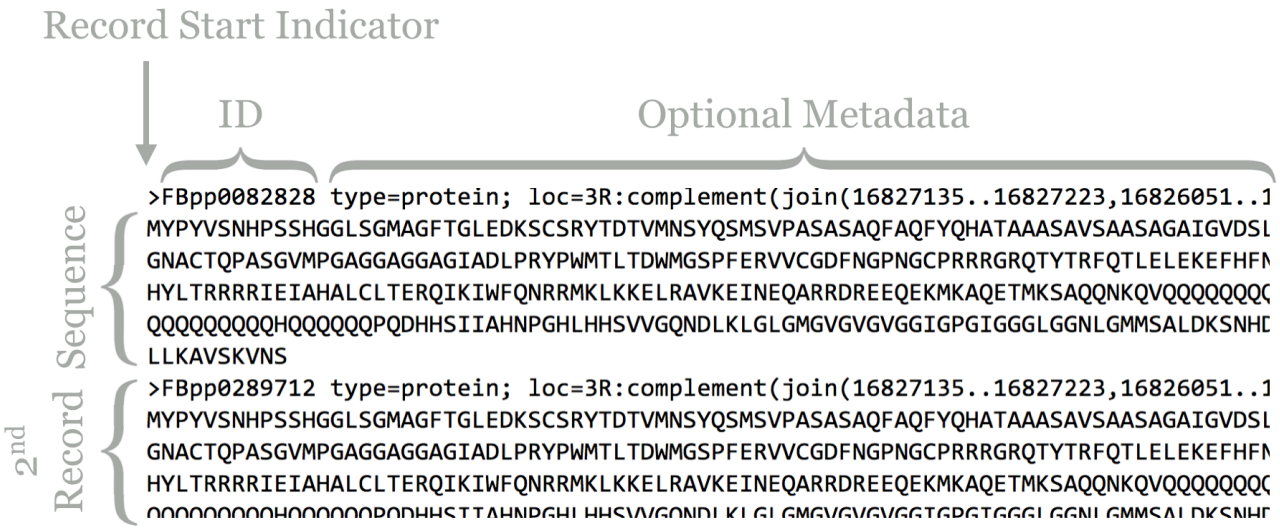
Download muscle binary

```
cd downloads

wget
http://www.drive5.com/muscle/downloads3.8.31/muscle3
.8.31_i86linux64.tar.gz -O muscle.tar.gz

gunzip muscle.tar.gz
tar -xf muscle.tar
ls
cp muscle3.8.31_i86linux64 $HOME/local/bin/muscle

# Needed to recognize new copy of muscle in $path
# Re-logout/login in or run tcsh's rehash
rehash
```

Oregon State UNIVERSITY

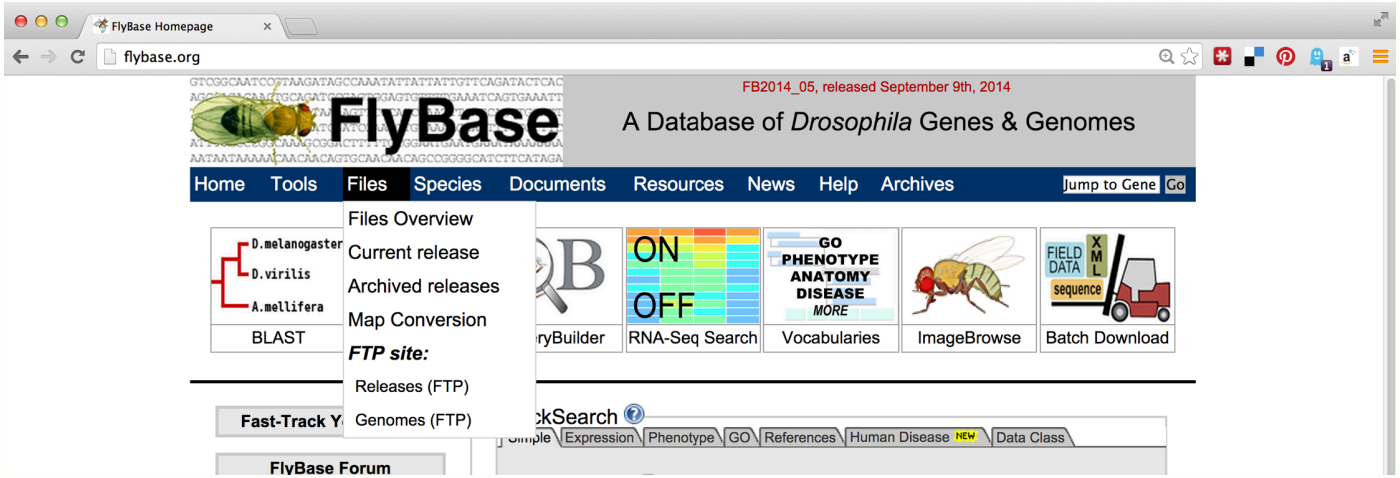
FASTA Format

- Goal: Find P450-1A1-like proteins in the Drosophila Melanogaster protein set.
- Require: Public datasets in **FASTA** format.
- ID and Metadata are not standardized.



Download protein set

- D. melanogaster protein set from flybase.org
- Can we identify, using **HMMER** and **muscle**, homologues (*same structural features but possibly not function*) of P450-1A1 genes in the Drosophila melanogaster protein dataset



Navigate to latest release

- "Downloads" / "Genomes (FTP)"
- "Drosophila_melanogaster"
- "dmel_r6.22_FB2018_03 "
- "fasta"
  - # gff and gtf hold gene location and function annotations
- "dmel-all-translation-r6.22.fasta.gz"

Copy URL of FASTA file

Oregon State UNIVERSITY

Download release

```
cd $HOME/projects
mkdir dmel_p450s
cd dmel_p450s
# Not using -O with wget
wget 'http://dmel … .tar.gz'
du -hs *gz          # Note size
gunzip *gz          # Uncompress
du -sh *fasta       # Increased?
less -S *fasta      # ID/Metadata
```

Oregon State UNIVERSITY

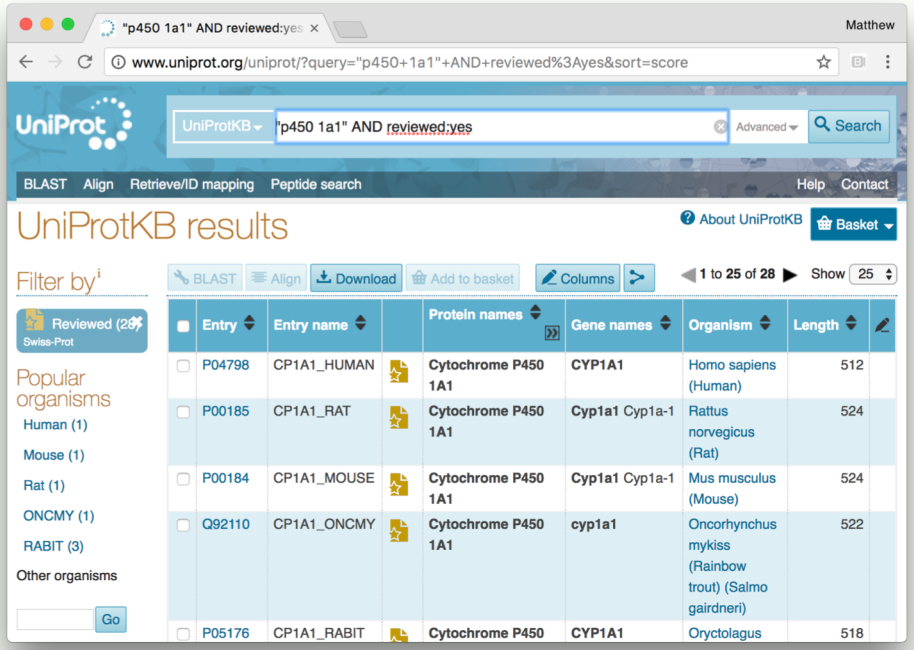
Uniprot.org

- **Uniprot.org** is a well-known protein database:
  - **TrEMBL** database
    - Many annotations have been assigned by automated homology searches and not reviewed.
  - **Swiss-Prot** (subset of TrEMBL)
    - Contains only sequences whose annotations have been manually reviewed.

Oregon State UNIVERSITY

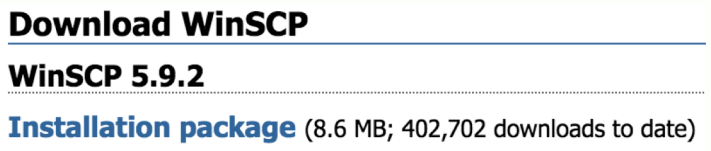
P4501A1 Genes (several species)

- Search: `"p450 1a1" AND reviewed:yes`
- "Download (all 28)" to your desktop
  - **`wget`** not easily done.

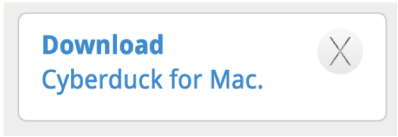


Upload to projects/dmel_p450s

- Use a local **SFTP** client (Secure File Transfer Protocol) to upload your **FASTA** file to crick
- SFTP Clients (OSU Recommended)
  - Windows
    - WinSCP (https://winscp.net/eng/download.php)
    - Download link under "Download WinSCP"



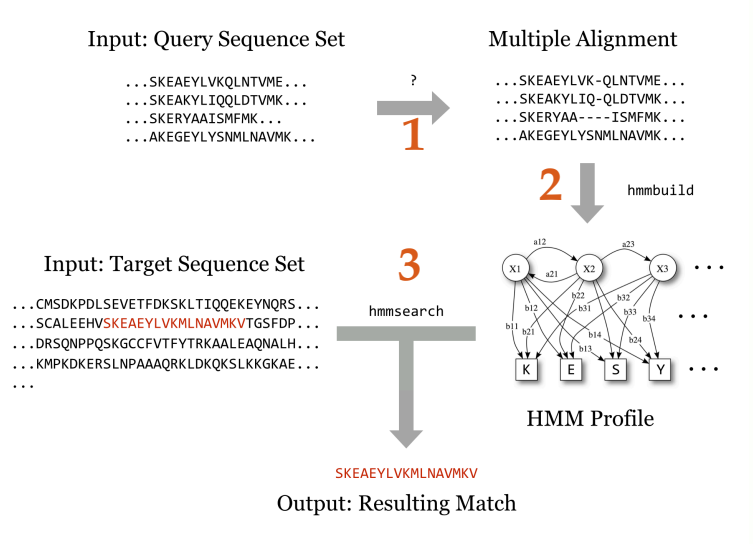
  - Mac
    - CyberDuck (https://cyberduck.io)
    - Download link on left (under duck), not at top.



Oregon State UNIVERSITY

Putting it all together

```
cd $HOME/projects/
cd dmel_450ps
ls # 2 FASTA files
where muscle
where hmmbuild
where hmmsearch
```



Step 1) Create a Multiple Sequence Alignment (MSA) of P450 sequences with **muscle**

Step 2) Build an HMM from this alignment with **hmmbuild**

Step 3) Search this HMM against the D.mel proteins with **hmmsearch**

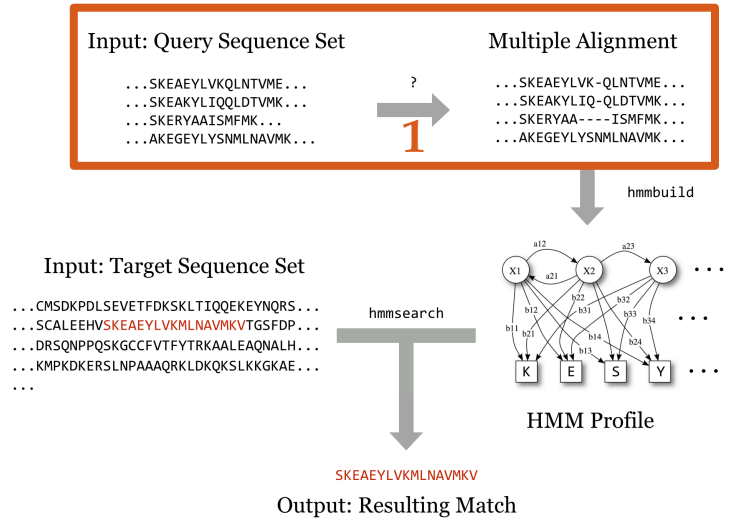
Oregon State University

Step 1: MSA of p450s with muscle

```
muscle -h # -help --help
```

Usage reported as:

```
muscle -in <inputfile>
       -out <outputfile>
```




```
muscle -in p450s.fasta -out p450s_aligned.fasta
```

Output is a FASTA file with alignments "padded"

```
less -S p450s_aligned.fasta
```

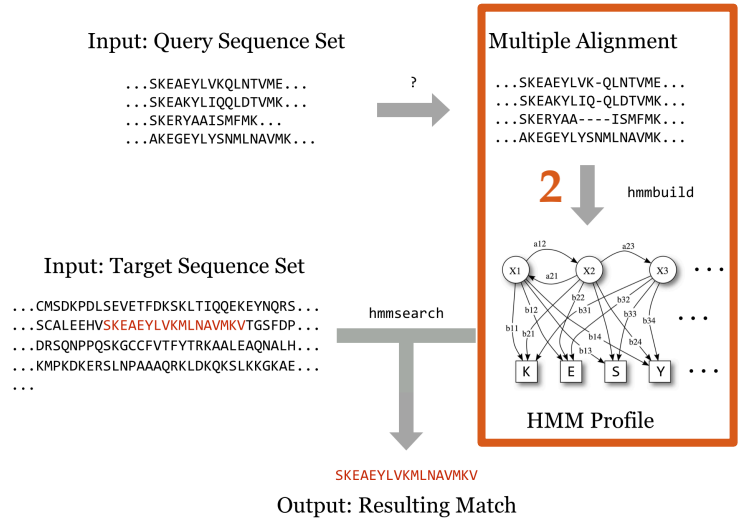
Oregon State University

Step 2: Build HMM from MSA

hmmbuild # hmmbuild -h

Usage reported as:

```
hmmbuild [-options]
         <hmmfile_out>
         <msafile>
```



[] optional, <> required, no –in / –out like **muscle**

```
hmmbuild p450s_aligned.hmm p450s_aligned.fasta
```

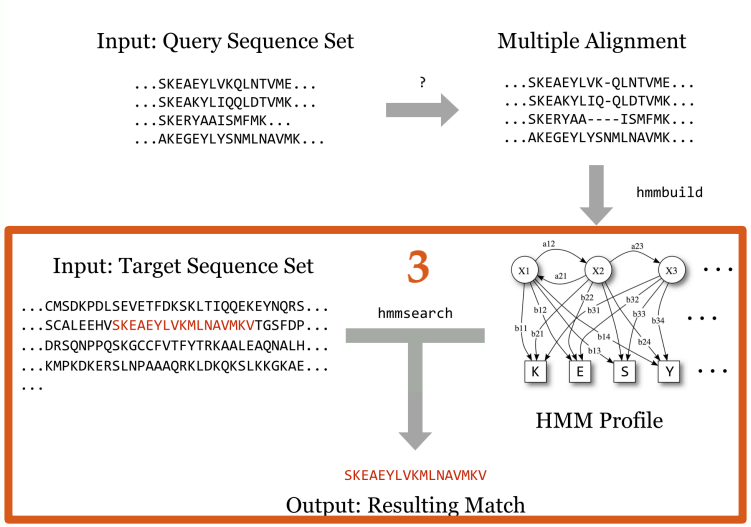
Output in text format; extensions arbitrary

```
less -S p450s_aligned.hmm
```

Oregon State University

Step 3: Search HMM vs. proteins

`hmmsearch # hmmsearch -h`



Usage reported as:

```
hmmsearch [options]
          <hmmfile>
          <seqdb>
```

Documentation specifies <seqdb> types, e.g., FASTA

```
hmmsearch p450s_aligned.hmm \
          dmel-all-translation-r6.22.fasta
```

Output to terminal (**stdout**), not to a file

```
ls
```

Oregon State University

stdout Standard Out

- **stdout** is an output stream that by default is printed to the terminal
- You can redirect the **stdout** output stream to a file via the **>** operator

```
hmmsearch p450s_aligned.hmm \
dmel-all-translation-r6.22.fasta \
> p450s_hmmsearch_dmel.txt
less -S p450s_hmmsearch_dmel.txt
```

- We will be covering redirection in far greater detail in the following classes…

Oregon State UNIVERSITY

Automation via shell script

```
nano runhummer.sh
```

```
#!/bin/tcsh
muscle -in p450s.fasta -out p450s_aligned.fasta
hmmbuild p450s_aligned.hmm p450s_aligned.fasta
hmmsearch p450s_aligned.hmm dmel-all-translation-r6.22.fasta > p450s_hmmsearch_dmel.txt
```

Oregon State UNIVERSITY

Environment variables

```
nano example.sh
```

```
#!/bin/tcsh
echo "first parameter is $1"
echo "second parameter is $2"
```

```
chmod +x ./example.sh
./example bob jones
```

Oregon State UNIVERSITY

Script with environment variables

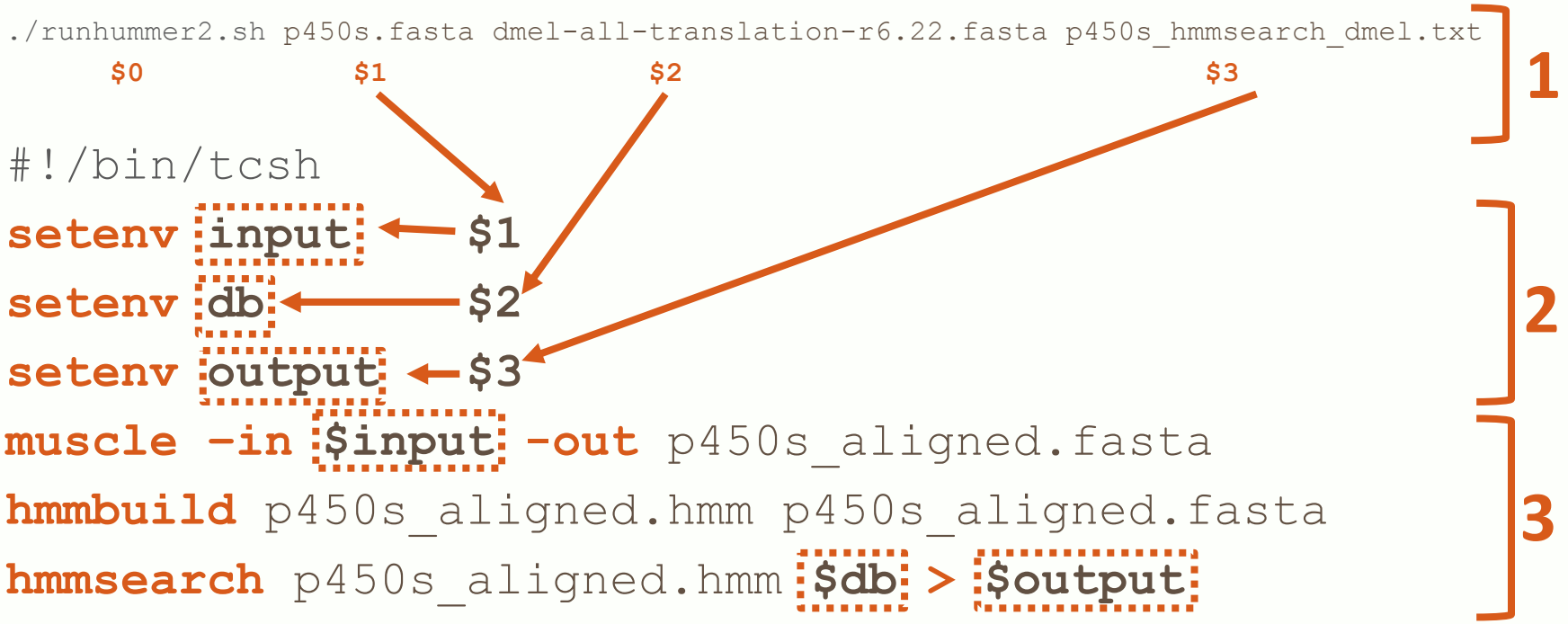
nano runhummer**2**.sh

```
#!/bin/tcsh
setenv input $1
setenv db $2
setenv output $3
muscle -in $input -out p450s_aligned.fasta
hmmbuild p450s_aligned.hmm p450s_aligned.fasta
hmmsearch p450s_aligned.hmm $db > $output
```

```
./runhummer2.sh p450s.fasta dmel-all-translation-r6.22.fasta p450s_hmmsearch_dmel.txt
      $0            $1                  $2                                $3
```

Run the above script providing 3 arguments (parameters) at the command line

Oregon State UNIVERSITY



1) Command line arguments **$1**, **$2**, **$3** are assigned (by the tcsh shell) the values p450s.fasta, dmel-all-translation-r6.22.fasta, and p450s_hmmsearch_dmel.txt respectively. **$0** is assigned the value of the program/command name, i.e., runhummer2.sh

2) User-defined variables **input**, **db**, and **output** are assigned the values of **$1**, **$2**, **$3** respectively via **setenv**

3) The value of the variable **$input** is used by **muscle** and the values of **$db** and **$ouput** are used by **hmmsearch**

Oregon State University

Command / Concept Review

- `uname -a`
- `rehash`
- `>`
- `$1, $2, etc.`
  `Environment variables`

`FASTA`

`SFTP`

`redirection`

`stdout`

Oregon State UNIVERSITY