

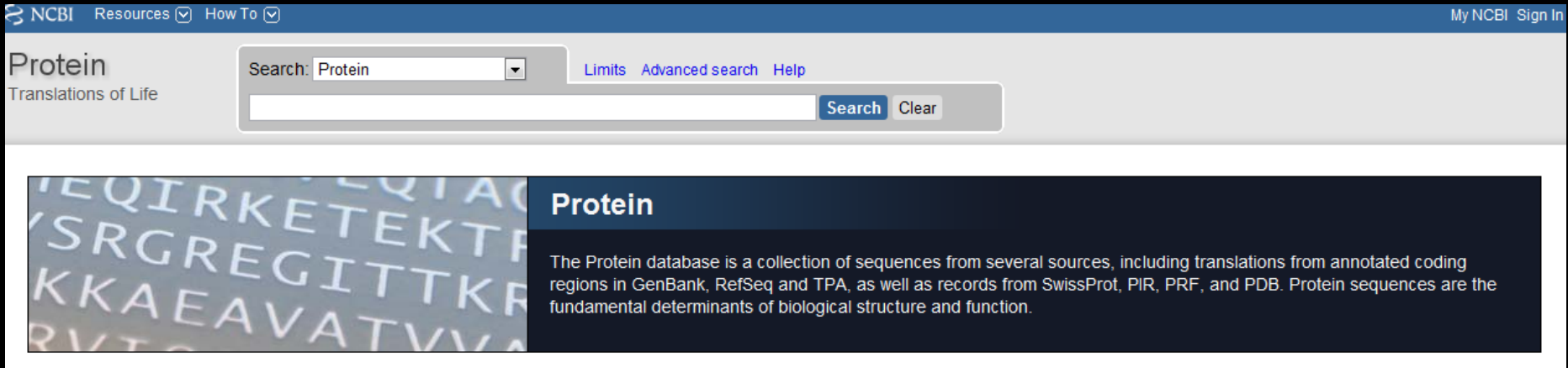
PHYLIP and phylogenies

New tools for Project 3

- The NCBI proteins database
 - Your resource for protein sequences, searchable by name and sequence.
- PHYLIP, a toolbox for generating and processing phylogenies
 - Neighbor
 - A simple phylogeny generator
 - Protpar
 - A maximum parsimony phylogeny generator
- Clustal2Matrix.py
 - A translator for ClustalW data (not ready yet)
- ClustalW

NCBI Protein database

<http://www.ncbi.nlm.nih.gov/protein/>



NCBI Resources ▾ How To ▾ My NCBI Sign In

Protein
Translations of Life

Search: Protein ▾ Limits Advanced search Help

Search Clear

Protein

The Protein database is a collection of sequences from several sources, including translations from annotated coding regions in GenBank, RefSeq and TPA, as well as records from SwissProt, PIR, PRF, and PDB. Protein sequences are the fundamental determinants of biological structure and function.

- Make sure the pulldown menu next to “Search:” says “Protein”
- Type in the search term you wish to look for

What you get from the Protein database

NCBI Resources ▾ How To ▾ My NCBI Sign In

Protein
Translations of Life

Search: Protein ▾ Save search Limits Advanced search Help

Beta-Lactamase Search Clear

- Here I am searching for Beta-lactamases, a classic protein for antibiotic resistance in bacteria

This search in Gene shows [26944 results](#), including:

- [bla1](#) (*Bacillus anthracis str. Ames*): beta-lactamase
- [CBU_0807](#) (*Coxiella burnetii* RSA 493): putative beta-lactamase
- [ampC](#) (*Shewanella oneidensis* MR-1): beta-lactamase

Results: 1 to 20 of 87107

- ☐ [beta-lactamase \[Shewanella baltica O6155\]](#)
1. 394 aa protein
Accession: YP_001050456.1 GI: 126174307
[GenPept](#) [FASTA](#) [Graphics](#) [Related Sequences](#) [Identical Proteins](#)
- ☐ [beta-lactamase \[Sinorhizobium medicae WSM419\]](#)
2. 386 aa protein
Accession: YP_001313637.1 GI: 150377041
[GenPept](#) [FASTA](#) [Graphics](#) [Related Sequences](#) [Identical Proteins](#)
- ☐ [beta-lactamase \[Paracoccus denitrificans PD1222\]](#)
3. 401 aa protein
Accession: YP_917217.1 GI: 119386162
[GenPept](#) [FASTA](#) [Graphics](#) [Related Sequences](#) [Identical Proteins](#)
- ☐ [beta-lactamase \[Cloning vector pBR322\]](#)
4. 286 aa protein
Accession: AAB59737.1 GI: 455370

- The first thing you notice:
 - There are a huge number of matches

Expect thousands and tens of thousands of matches.

- Applications projects: you need a large database like this to get enough diversity

What you get from the Protein database

NCBI Resources How To My NCBI Sign In

Protein
Translations of Life

Search: Protein Save search Limits Advanced search Help

Beta-Lactamase Search Clear

- Here I am searching for Beta-lactamases, a classic protein for antibiotic resistance in bacteria

This search in Gene shows [26944 results](#), including:

[bla1](#) (*Bacillus anthracis str. Ames*): beta-lactamase
[CBU_0807](#) (*Coxiella burnetii* RSA 493): putative beta-lactamase
[ampC](#) (*Shewanella oneidensis* MR-1): beta-lactamase

Results 1 to 20 of 27407

☐ [beta-lactamase \[Shewanella baltica OS155\]](#)
394 aa protein
Accession: YP_001050456.1 GI: 126174307
[GenPept](#) [FASTA](#) [Graphics](#) [Related Sequences](#) [Identical Proteins](#)

☐ [beta-lactamase \[Sinorhizobium medicae WSM419\]](#)
386 aa protein
Accession: YP_001313637.1 GI: 150377041
[GenPept](#) [FASTA](#) [Graphics](#) [Related Sequences](#) [Identical Proteins](#)

☐ [beta-lactamase \[Paracoccus denitrificans PD1222\]](#)
401 aa protein
Accession: YP_917217.1 GI: 119386162
[GenPept](#) [FASTA](#) [Graphics](#) [Related Sequences](#) [Identical Proteins](#)

☐ [beta-lactamase \[Cloning vector pBR322\]](#)
286 aa protein
Accession: AAB59737.1 GI: 455370

- Each entry points to a single protein sequence
- Immediately next to each protein name is the source organism
- Shewanella Baltica is a bacterium that is able to alter metals so that they can be better broken down in the environment

What you get from the Protein database

NCBI Resources ▾ How To ▾ My NCBI Sign In

Protein
Translations of Life

Search: Protein ▾ Save search Limits Advanced search Help

Beta-Lactamase Search Clear

- Here I am searching for Beta-lactamases, a classic protein for antibiotic resistance in bacteria

This search in Gene shows [26944 results](#), including:

[bla1](#) (*Bacillus anthracis* str. Ames): beta-lactamase
[CBU_0807](#) (*Coxiella burnetii* RSA 493): putative beta-lactamase
[ampC](#) (*Shewanella oneidensis* MR-1): beta-lactamase

Results: 1 to 20 of 87107

1. [beta-lactamase \[Bacillus anthracis str. Ames\]](#)
394 aa protein
Accession: YP_001054432.1 GI: 1261174307
[GenPept](#) [FASTA](#) [Graphics](#) [Related Sequences](#) [Identical Proteins](#)

2. [beta-lactamase \[Sinorhizobium medicae WSM419\]](#)
386 aa protein
Accession: YP_001313637.1 GI: 150377041
[GenPept](#) [FASTA](#) [Graphics](#) [Related Sequences](#) [Identical Proteins](#)

3. [beta-lactamase \[Paracoccus denitrificans PD1222\]](#)
401 aa protein
Accession: YP_917217.1 GI: 119386162
[GenPept](#) [FASTA](#) [Graphics](#) [Related Sequences](#) [Identical Proteins](#)

4. [beta-lactamase \[Cloning vector pBR322\]](#)
286 aa protein
Accession: AAB59737.1 GI: 455370

- This is not an irrelevant detail
- The number of amino acids in the protein can be a very easy indicator that you can or cannot use this sequence
 - Most sequences of the same protein have similar lengths. Something very short or very long should not be used

Another piece of important information

Gene Information

This search in Gene shows [26944 results](#), including:

- [bla1](#) (*Bacillus anthracis* str. Ames): beta-lactamase
- [CBU_0807](#) (*Coxiella burnetii* RSA 493): putative beta-lactamase
- [ampC](#) (*Shewanella oneidensis* MR-1): beta-lactamase

Gene

Results: 1 to 20 of 87107 Page [1](#) of 4356 [Next >](#) [Last >>](#)

- ☐ [beta-lactamase \[Shewanella baltica OS155\]](#)
1. 394 aa protein
Accession: YP_001050456.1 GI: 126174307
[GenPept](#) [FASTA](#) [Graphics](#) [Related Sequences](#) [Identical Proteins](#)
- ☐ [beta-lactamase \[Sinorhizobium medicae WSM419\]](#)
2. 386 aa protein
Accession: YP_001313637.1 GI: 150377041
[GenPept](#) [FASTA](#) [Graphics](#) [Related Sequences](#) [Identical Proteins](#)
- ☐ [beta-lactamase \[Paracoccus denitrificans PD1222\]](#)
3. 401 aa protein
Accession: YP_917217.1 GI: 119386162
[GenPept](#) [FASTA](#) [Graphics](#) [Related Sequences](#) [Identical Proteins](#)
- ☐ [beta-lactamase \[Cloning vector pBR322\]](#)
4. 286 aa protein
Accession: AAB59737.1 GI: 455370

Top Organisms [Tree]

- Escherichia coli (13228)
- Escherichia coli K-12 (8599)
- Escherichia coli str. K-12 substr. W3110 (8453)
- Bacillus cereus (3422)
- Salmonella enterica (1996)
- All other taxa (75748)
- [More...](#)

Find related

Database: [Find items](#)

Search data

[All Fields](#)

- Applications projects: you will want broad representation of many organisms. You can use this filter to get representatives from the largest groups
- Make sure to weed out similar groups, like the top two here

Tree view

Gene Information

All (87107)

[Bacteria \(79326\)](#)

[Related Structures \(77115\)](#)

[RefSeq \(37734\)](#)

[Manage Filters](#)

Page 1 of 4356 [Next >](#) [Last >>](#)

▼ Top Organisms [\[Tree\]](#)

- Escherichia coli (13,281)
- Escherichia coli K-12 (8599)
- Escherichia coli str. K-12 substr. W3110 (8453)
- Bacillus cereus (3422)
- Salmonella enterica (1996)
- All other taxa (75748)
- [More...](#)

Find related data

Database:

[Find items](#)

Search details

Beta-Lactamase[All Fields]

Send to: ☐

Filter your results:

All (87107)

[Bacteria \(79326\)](#)

[Related Structures \(77115\)](#)

[RefSeq \(37734\)](#)

[RefSeq](#) [Filters](#)

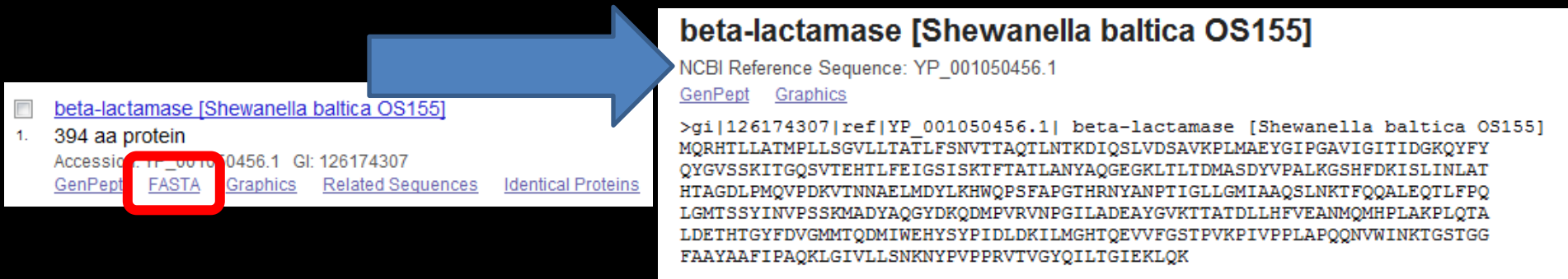
Page 1 of 4356 [Next >](#) [Last >>](#)

▼ Taxonomic Groups [\[List\]](#)

- bacteria (79326)
 - proteobacteria (43985)
 - g-proteobacteria (29533)
 - a-proteobacteria (6854)
 - b-proteobacteria (5840)
 - more... (1758)
 - firmicutes (19235)
 - Bacillales (10445)
 - more... (8790)
 - actinobacteria (8340)
 - Actinomycetales (7871)
 - more... (469)
 - CFB group bacteria (3168)
 - cyanobacteria (1085)
 - more... (3513)
- eukaryotes (3034)
 - fungi (1326)
 - animals (918)
 - more... (790)
- other sequences (2221)
- archaea (1936)
 - euryarchaeotes (1202)
 - crenarchaeotes (667)
 - more... (67)
- viruses (106)
- unclassified (66)

- Applications project: This view provides a sense of how closely related the different organisms are. Selecting sequences carefully from groups will help you pick a better representative set.

Getting data out of the Protein database



The screenshot shows the NCBI protein database entry for **beta-lactamase [Shewanella baltica OS155]**. On the left, a summary box lists the protein as 394 aa, with accession number YP_001050456.1 and GI: 126174307. A red box highlights the [FASTA](#) link. A large blue arrow points from this link to the right, where the full protein sequence is displayed. The sequence header is `>gi|126174307|ref|YP_001050456.1| beta-lactamase [Shewanella baltica OS155]`, followed by the amino acid sequence.

beta-lactamase [Shewanella baltica OS155]
NCBI Reference Sequence: YP_001050456.1
[GenPept](#) [Graphics](#)
>gi|126174307|ref|YP_001050456.1| beta-lactamase [Shewanella baltica OS155]
MQRHTLLATMPLLSGVLLTATLFSNVTTAQTILNTKDIQSLVDSAVKPLMAEYGIPGAVIGITIDGKQYFY
QYGVSSKITGQSVTEHTLFEIGSISKTFTATLANYAQQEGKLTLTDMASDYVPALKGSHFDKISLINLAT
HTAGDLPMQVPDKVTNNAEELMDYLKHWQPSFAPGTHRNYPNTIGLLGMIAAQSINKTFQQALETLPFPQ
LGMTSSYINVPSSKMADYAQQGYDKQDMPVRVNPGLADEAYGVKTTATDLLHFVEANMQMHPPLAKPLQTA
LDEHTGYFDVGMMTQDMIWEHYSYPIDLDKILMGHTQEVVFGSTPVKPIVPLAPQQNVWINKTGSTGG
FAAYAAFIPAQKLGIVLLSNKNYPVPPRVTVGYQILTGIEKLQK

- When you decide that you want to use a sequence in your project
 - Click on the FASTA link, and the Protein database provides a fasta entry
 - Cut and paste after the “>” symbol, to the end of the text.
- Note: you may want to modify the header that they give you to be something that you prefer
 - This header will help you find the same sequence again later, however

PHYLIP



- Famous and venerable phylogeny software
 - Developed for over 30 years by Joe Felsenstein (and his students) at the University of Washington
 - The baseline for phylogeny software
- Many components, but the two you will use are:
 - Neighbor
 - Protpar

Neighbor

- This software generates neighbor joining and UPGMA trees from distance matrices:

	A	B	C
A			
B	.3		
C	.5	.8	

A simple distance matrix

Implementation Projects:

- This is a useful benchmark for testing your UPGMA tree generators

How to use neighbor

- First, copy this directory to your experiment dir:

```
cp -r /proj/cse308/Project3 /proj/cse308-  
<username>
```

- Then enter the directory:

```
cd /proj/cse308-<username>/Project3/Phylip
```

- Enter the command:

```
./neighbor
```

- This will start neighbor, and you will see this:

```
neighbor: can't find input file "infile"  
Please enter a new file name> ■
```

- Neighbor works much the way ClustalW works,
by taking stdin

Providing input for neighbor

```
neighbor: can't find input file "infile"  
Please enter a new file name> █
```

- Here type the following:

neighborExample.grid

- This is an input file that contains a distance matrix. We will explain what this is and how you get this soon.

A minor error you might see:

```
neighbor: can't find input file "infile"  
Please enter a new file name> neighborExample.grid  
  
neighbor: the file "outfile" that you wanted to  
use as output file already exists.  
Do you want to Replace it, Append to it,  
write to a new File, or Quit?  
(please type R, A, F, or Q)
```

- Sometimes one of the output files will already exist.
 - This is because neighbor defaults to generating output files with the same name: outfile.
 - If you don't care, simply press "R" to over-write it.
 - If you want to save your output file to a new name, press F to give it your own name

The main menu for neighbor

```
Neighbor-Joining/UPGMA method version 3.69

Settings for this run:
N      Neighbor-joining or UPGMA tree?  Neighbor-joining
O      Outgroup root?                  No, use as outgroup species  1
L      Lower-triangular data matrix?    No
R      Upper-triangular data matrix?    No
S      Subreplicates?                   No
J      Randomize input order of species? No. Use input order
M      Analyze multiple data sets?      No
0      Terminal type (IBM PC, ANSI, none)? ANSI
1      Print out the data at start of run No
2      Print indications of progress of run Yes
3      Print out tree                    Yes
4      Write out trees onto tree file?   Yes

Y to accept these or type the letter for one to change
```

- Here, press “N” to toggle Neighbor-joining to UPGMA
- Press “L” to toggle Lower-triangular data matrix to “On”
- Then press “Y” to accept the changes and run neighbor to generate the tree.

The menu, set correctly.

```
Neighbor-Joining/UPGMA method version 3.69

Settings for this run:
N      Neighbor-joining or UPGMA tree?  UPGMA
L      Lower-triangular data matrix?    Yes
R      Upper-triangular data matrix?    No
S      Subreplicates?                   No
J      Randomize input order of species? No. Use input order
M      Analyze multiple data sets?      No
0      Terminal type (IBM PC, ANSI, none)? ANSI
1      Print out the data at start of run No
2      Print indications of progress of run Yes
3      Print out tree                    Yes
4      Write out trees onto tree file?   Yes

Y to accept these or type the letter for one to change
Y
```

- This is what your menu should look like after you've updated the settings
- (this slide is partially for reference)
- You can ignore the information that neighbor outputs to the screen; it's not terribly useful for the project

Another error you might see

```
neighbor: the file "outtree" that you wanted to  
use as output tree file already exists.  
Do you want to Replace it, Append to it,  
write to a new File, or Quit?  
(please type R, A, F, or Q)
```

- The other output file, “outtree” also defaults to this name
 - Overwrite it by pressing “R”
 - Create a new file with a custom name by pressing “F”

What neighbor generates

- Lets look at the outfile first
- `less ./outfile`
- The outfile is a human readable file meant to give you an idea of the composition of the tree

```
less ./outfile
```

The other output file from neighbor

- This file is less useful for you initially, but will be useful for generating pretty trees

less ./outtree

- Here, the output is not designed to be human readable. This format is called Newick


```
[07:55 AM][chen@titania Phylip] less outtree
(((((((1auo:0.10043,1aur:0.10043):0.95379,2h1i:1.05422):0.25827,
(((1evq:0.28437,1qz3:0.28437):0.16165,(1u4n:0.26559,2hm7:0.26559):0.18044):0.42081,
(1jji:0.65050,2c7b:0.65050):0.21632):0.36467,(2o7r:0.13683,
2o7v:0.13683):1.09467):0.08099):0.25520,(1r1d:0.17612,1tqh:0.17612):1.39157):0.0424
,
117r:1.61014):0.09273,(((1k4y:0.36878,((1mx1:0.16483,1mx9:0.16483):0.08155,
(((1mx5:0.15597,(1ya4:0.14254,(1yah:0.10982,2dqy:0.10982):0.03272):0.01342):0.012
4,
(2dqz:0.14812,2dr0:0.14812):0.02028):0.01394,(2hrq:0.13125,
2hrr:0.13125):0.05110):0.03142,2h7c:0.21376):0.01341,(1ya8:0.20241,
1yaj:0.20241):0.02476):0.01921):0.12241):0.68322,((2jey:0.08334,
2jez:0.08334):0.02550,2jf0:0.10884):0.94316):0.09190,(2ogs:0.18988,
2ogt:0.18988):0.95403):0.17810,2fj0:1.32200):0.38086):0.11228,
(1M33:1.51737,2r11:1.51737):0.29778);
outtree (END)
```


Using Newick files

- Newick is a precise description of the tree, including edge lengths.
- If you want to render a pretty tree, you can give the renderer the newick file.
- One such piece of software is phylowidget:
<http://www.phylowidget.org>

PhyloWidget Quickstart

Note: a URL pointing to your favorite Newick or Nexus file is also valid input for PhyloWidget Full.

 **Open in PhyloWidget Full**

 **Open in PhyloWidget Lite**

Input that Neighbor takes

- Lets open up the input file for neighbor

```
less ./neighborExample.grid
```

[illegible]

- The input is a lower triangular grid of distances between identifiers (left column). The file starts with a number, indicating total # of indicators
- These numbers are artificial, but in your data will reflect sequence identities between protein sequences:

$$(100 - \text{seqId}) / 100$$

Protpars

- Protpars attempts to compute a maximum parsimony tree based on your input sequence.
- Lets look at the input first:

```
less protparsExample.txt
```

	5	10
Alpha	ABCDEF	GHIK
Beta	AB--	EFGHIK
Gamma	?BCD	SFG*??
Delta	CIKDE	F
Epsilon	DIKDE	F

- Here the first two numbers are the number of sequences, and the number of nucleotides
- You can see that protpars can run this computation using wildcards as well: ?, *

Running protpars

- Run protpars with this simple command

`./protpars`

- Protpars will give you a similar error as neighbor:

```
protpars: can't find input file "infile"  
Please enter a new file name>
```

- Here, type in the other sample input file:

`protparsExample.txt`

- This is the input example we just saw

Protpars Errors you might see

- This sort of file error can happen with protpars as well as neighbor

```
protpars: can't find input file "infile"  
Please enter a new file name> protparsExample.txt  
  
protpars: the file "outfile" that you wanted to  
          use as output file already exists.  
          Do you want to Replace it, Append to it,  
          write to a new File, or Quit?  
          (please type R, A, F, or Q)
```

- Just replace (“R”) the file with another outfile, for this example
- When you are running your own data, you will probably want to create a new file, using “F”

Options for Protpars

```
Protein parsimony algorithm, version 3.69

Setting for this run:
U      Search for best tree?      Yes
J      Randomize input order of sequences?  No. Use input order
O      Outgroup root?            No, use as outgroup species  1
T      Use Threshold parsimony?   No, use ordinary parsimony
C      Use which genetic code?    Universal
W      Sites weighted?           No
M      Analyze multiple data sets? No
I      Input sequences interleaved? Yes
0      Terminal type (IBM PC, ANSI, none)? ANSI
1      Print out the data at start of run  No
2      Print indications of progress of run Yes
3      Print out tree              Yes
4      Print out steps in each site  No
5      Print sequences at all nodes of tree No
6      Write out trees onto tree file? Yes

Are these settings correct? (type Y or the letter for one to change)
_
```

- Press “U” to search for the best tree.
- Press “J” to randomize the input order
- Then press “Y” to start the run
- Again you might see this error:, just press “R”

```
protpars: the file "outtree" that you wanted to
use as output tree file already exists.
Do you want to Replace it, Append to it,
write to a new File, or Quit?
_ (please type R, A, F, or Q)
```

Outputs for Protpars

- Protpars also generates an outfile and an outtree.
- The outtree is a Nevick file exactly like the output from neighbor
- Lets look at the the outfile:

less ./outfile

- Here Protpars finds three possible high-parsimony trees

```
Protein parsimony algorithm, version 3.69
```

```
3 trees in all found
```

```

+-----Gamma
!
+--2      +--Epsilon
! ! +--4
! +--3 +--Delta
1      !
!      +-----Beta
!
+-----Alpha

remember: this is an unrooted tree!
```

```
requires a total of      16.000
```

```

+--Epsilon
+--4
+--3 +--Delta
! !
+--2 +-----Gamma
! !
1 +-----Beta
!
+-----Alpha

remember: this is an unrooted tree!
```

```
requires a total of      16.000
```

```

+--Epsilon
+-----4
!      +--Delta
+--3
! !      +--Gamma
1 +-----2
!      +--Beta
!
+-----Alpha

remember: this is an unrooted tree!
```

How you can use Protpars

- Applications projects:
 - When you have assembled your multiple sequence alignments, you should feed these alignments to protpars to see how diverse your tree is
 - A diverse tree will be evenly branched, rather than having lots of nested trees in one branch.
- Implementations projects:
 - You can use Protpars as a sort of answer key, to see if your maximum parsimony algorithms are moving in the same directory
 - Computing maximum parsimony is NP-hard, so there is no right answer that can be easily computed.

Clustal2Matrix.py

- This python script will take clustalW output (pairwise sequence identities) and use it to generate the distance matrix used for running neighbor.
- This is so that you don't have to do this processing
- I am still coding this, it will be ready this weekend.

Questions