NCBI FTP (File Transfer Protocol) Server

The NCBI FTP server allows an access to most of the data represented at the webpages, but in a structured way, which can be easily accessed in an automatic process. In contrast to the SFTP server, at least the NCBI FTP server does not require an account. The login will be with the user “anonymous” (without any password) automatically.

First you need to login to Alex’s account via ssh (Material 2).
Change the directory to ~/YOUR_NAME/.

We will downloads the *N. viennensis* genome and the respective feature table as an example:

```
wget
ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/698/785/GCA_000698785.1_ASM69878v1/
GCA_000698785.1_ASM69878v1_protein.faa.gz

wget
ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/698/785/GCA_000698785.1_ASM69878v1/
GCA_000698785.1_ASM69878v1_feature_table.txt.gz
```

Uncompress both files:

```
gunzip GCA_000698785.1_ASM69878v1_protein.faa.gz
gunzip GCA_000698785.1_ASM69878v1_feature_table.txt.gz
```

This command searches for all amo genes, extxracts the accession numbers and saves them in a file:

```
grep amo[A-Z] GCA_000698785.1_ASM69878v1_feature_table.txt|grep CDS |cut -f 11 >
list_amo_IDS.txt
```

All steps are connected with the pipe ‘|’. You can also run them separately to understand the function of each step:

```
grep amo[A-Z] GCA_000698785.1_ASM69878v1_feature_table.txt
grep CDS GCA_000698785.1_ASM69878v1_feature_table.txt
cut -f 11 GCA_000698785.1_ASM69878v1_feature_table.txt
```

Now we want to extract the amo sequences from the genome. First, we need to load the blast progamm:

```
module load ncbiblastplus/2.6.0
```

After that, we need to create an indexed database:

```
makeblastdb -dbtype prot -in GCA_000698785.1_ASM69878v1_protein.faa -parse_seqids
```

There are three ways to extract sequences based on accession numbers.

This command extracts one sequence (amoA):

```
blastdbcmd -db GCA_000698785.1_ASM69878v1_protein.faa -entry AIC17003.1
```

This command extracts the sequences amoA, amoB and amoX:

```
blastdbcmd -db GCA_000698785.1_ASM69878v1_protein.faa –entry
AIC17003.1,AIC17117.1,AIC17004.1
```

This command extracts all amo sequences basend on the accession number saved in the file list_amo_IDS.txt:

```
blastdbcmd -db GCA_000698785.1_ASM69878v1_protein.faa -entry_batch
list_amo_IDS.txt
```