Chapter 4 in October

There is no “code” in Chapter 4 – but the HW problem is big.

Please read book = reflection this week on reading!

Student Learning Outcomes:

1) Guessing the function of a protein using homologs

2) Identifying likely genes in a DNA sequence

3) IDE = interactive development environment

4) Working together with a team to build a larger piece of code

Your mid-term project:

```
def geneFinder(DNA,minLen):
    """Find genes in a piece of DNA. Return list of start, end
    protein, sorted by start coordinate."""
```

Your mentor will break your team into 3-4 mini teams. Each team will submit their code as a group, but individual "reports" are required.

geneFinder will call many other functions:
Each team member will select certain functions to write. Then these functions will be combined together for the final code.

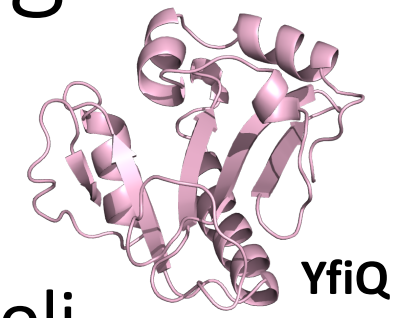
I'm not 100% sure when the code and report will be due, but sometime between Oct 25 – Nov 2. Full instructions will be released by Oct 12.

Your Team's mid-term project

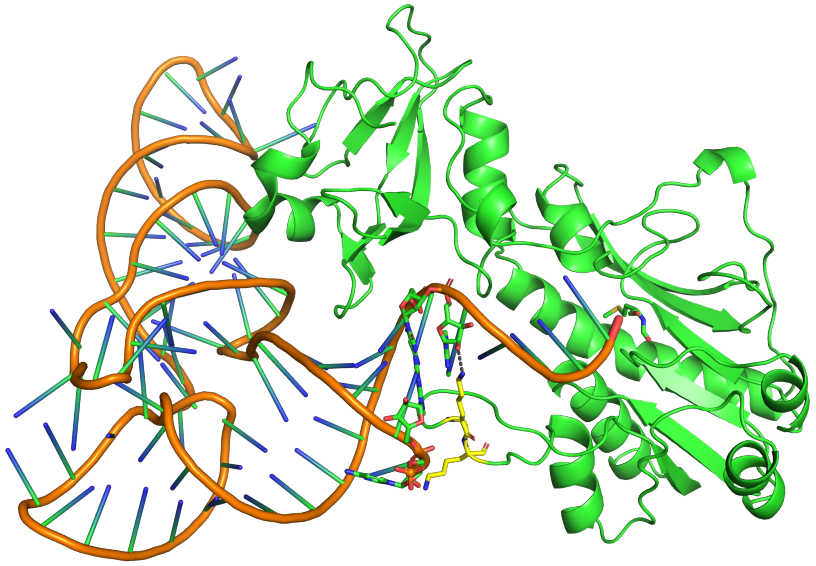
The famous biologist, Professor P.I. Pette, has been studying Salmonella in her lab. She has obtained the following sequence from a novel pathogenic strain: salDNA.fa. Prof. Pette obtained this from a region unique to Salmonella. She believes it is involved in pathogenisis, and may be from pathogenicity island SPI1.
Your project is to write a code/function called GeneFinder.py which will help Prof Pette determine the function of this sequence.

To do this:
A) Identify any genes
B) Blast the genes at NCBI BLAST; using the links which NCBI BLAST gives you determine the name and function of your mystery gene(s).
C) Write a report explaining what you have found.

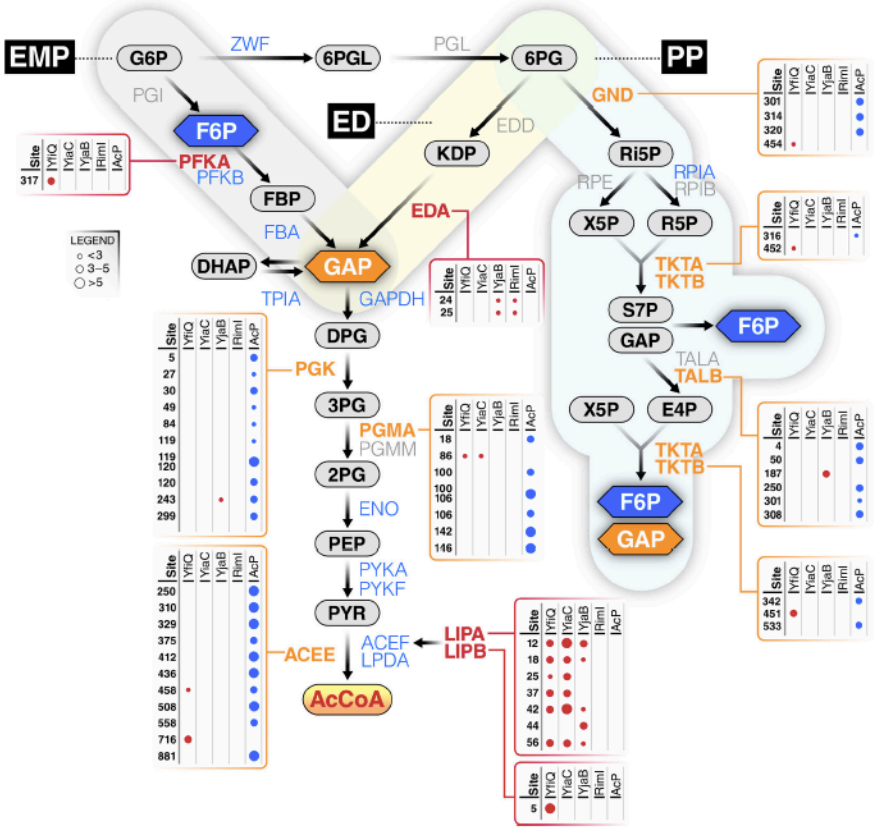
Use NCBI BLAST to find homologs



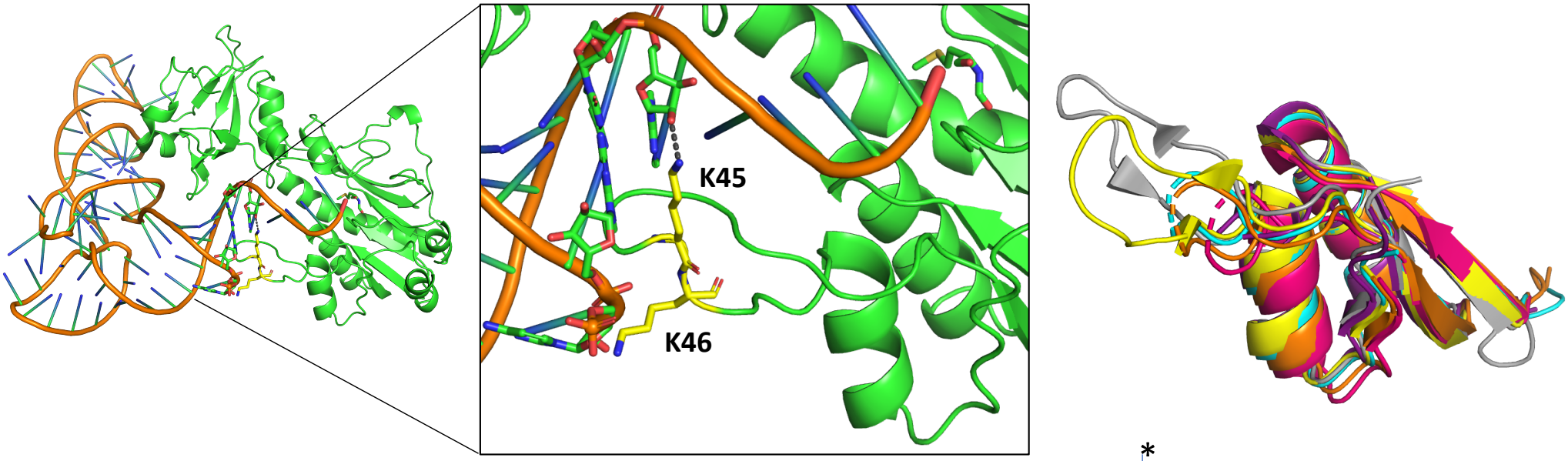
- With Chemistry Professor Misty Kuhn, we are studying K acetylation enzymes in Ecoli
- Proteins involved in purine metabolism may be acetylated



Protein 2fmt



Find homologs to determine conservation of active site amino acids



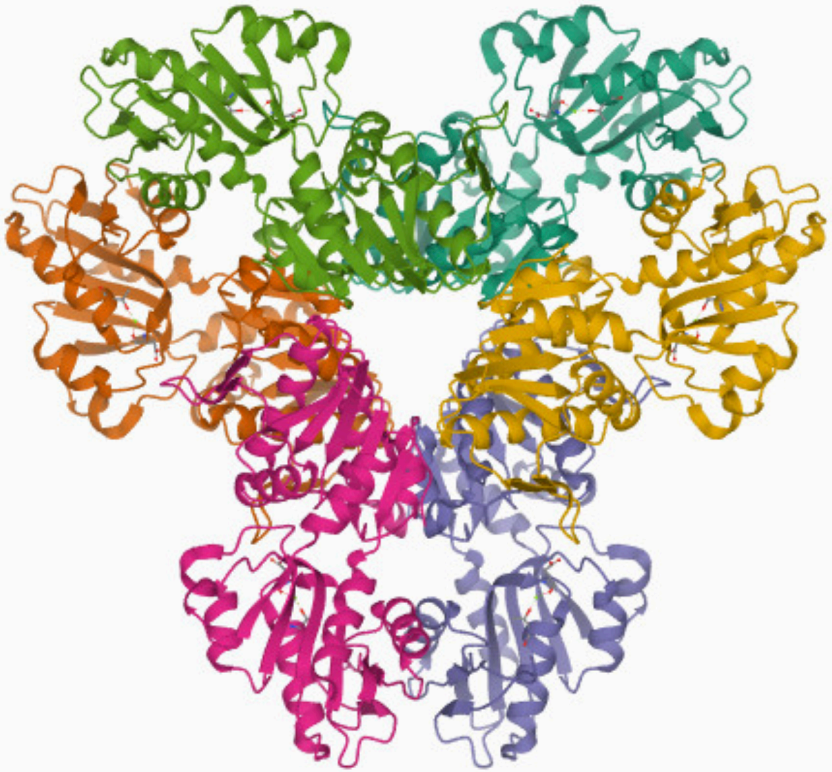
H-bonds with 2’-OH of adenosine on tRNA in 2fmt structure

```
                1         10        20        30        40        50        60
E_coli          MSESLRIIFAGTPDFAARHLDALLSSGHNVVGVFTQPDRPAGRGKKLMPSPVKVLAEEKG
V_cholerae      MSQSLRIVFAGTPDFAARHLAALLSSEHEIIAVYTQPERPAGRGKKLTASPVKTLALEHN
Y_pestis        MSDSLRIIFAGTPDFAARHLGALLSSQHKIVGVFTQPDRPAGRGNKLTPSPVKILAEHHG
P_aeruginosa    MSQALRIVFAGTPEFAAEHLKALLDTPHRIVAVYTQPDRPAGRGQKLMPSAVKSLALEHG
C_burnetii      ..MSLKIVFAGTPQFAVPTLRALIDSSHRVLAVYTQPDRPSGRGQKIMESPVKEIARQNE
B_anthracis     ...MIKVVFMGTPDFSVPVLRRLIEDGYDVIGVVTQPDRPVGRKKVLTPTPVKVEAEKHG
```

K mostly conserved!

BLAST finds homologs in Shingella!

Biological assembly of E Coli's
Phosphorybosylpyrophosphate synthetase



- Gene = prs
- UNIPROT = P0A717
- Protein Data Base ID = 4S2U
- Asymmetric Unit =

MPDMKLFAGNATPELAQRIANRLYTSLGDAAVG
RFSDGEVSVQINENVRGGDIFIIQSTCAPTNDNL
MELVVMVDALRRASAGRITAVIPYFGYARQDRRV
RSARVPITAKVVADFLSSVGVDRVLTVDLHAEQIQ
GFFDVPVDNVFGSPILLEDMLQLNLDNPIVVSPDI
GGVVRARAIAKLLNDTDMAIIDKRRPRANVSQV
MHIIGDVAGRDCVLVDDMIDTGGTLCKAAEALK
ERGAKRVFAYATHPIFSGNAANNLRNSVIDEVVV
CDTIPLSDEIKSLPNVRTLTLSGMLAEAIRRISNEES
ISAMFEH

Enter Sequence in Box

Standard Protein BLAST

blastn | blastp | blastx | tblastn | tblastx

BLASTP programs search protein databases using a protein query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

Clear

Query subrange

From

To

```
>unnamed protein product
MPDMKLFAGNATPELAQRIANRLYTSLGDAAVGRFSDGEVSVQINENVRGGDIFIIQSTCAPTNDNL
MELVVMVDALRRASAGRITAVIPYFGYARQDRRVRSARVPITAKVVADFLSSVGVDRVLTVDLHAEQ
IQGFFDVPVDNVFGSPILLEDMLQLNLDNPIVVSPDIGGVVRARAIAKLLNDTDMAIIDKRRPRANV
SQVMHIIGDVAGRDCVLVDDMIDTGGTLCKAAEALKERGAKRVFAYATHPIFSGNAANNLRNSVIDE
```

Or, upload file — Choose File — No file chosen

Job Title — Protein Sequence

Enter a descriptive title for your BLAST search

Align two or more sequences

BLAST has New Default Parameters and Search Limits.

Click here for more info.

Choose Search Set

Database — Non-redundant protein sequences (nr)

Organism (Optional) — exclude +

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude (Optional) — Models (XM/XP) | Non-redundant RefSeq proteins (WP) | Uncultured/environmental sample sequences

Program Selection

Algorithm

- Quick BLASTP (Accelerated protein-protein BLAST)
- blastp (protein-protein BLAST)
- PSI-BLAST (Position-Specific Iterated BLAST)
- PHI-BLAST (Pattern Hit Initiated BLAST)
- DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

BLAST

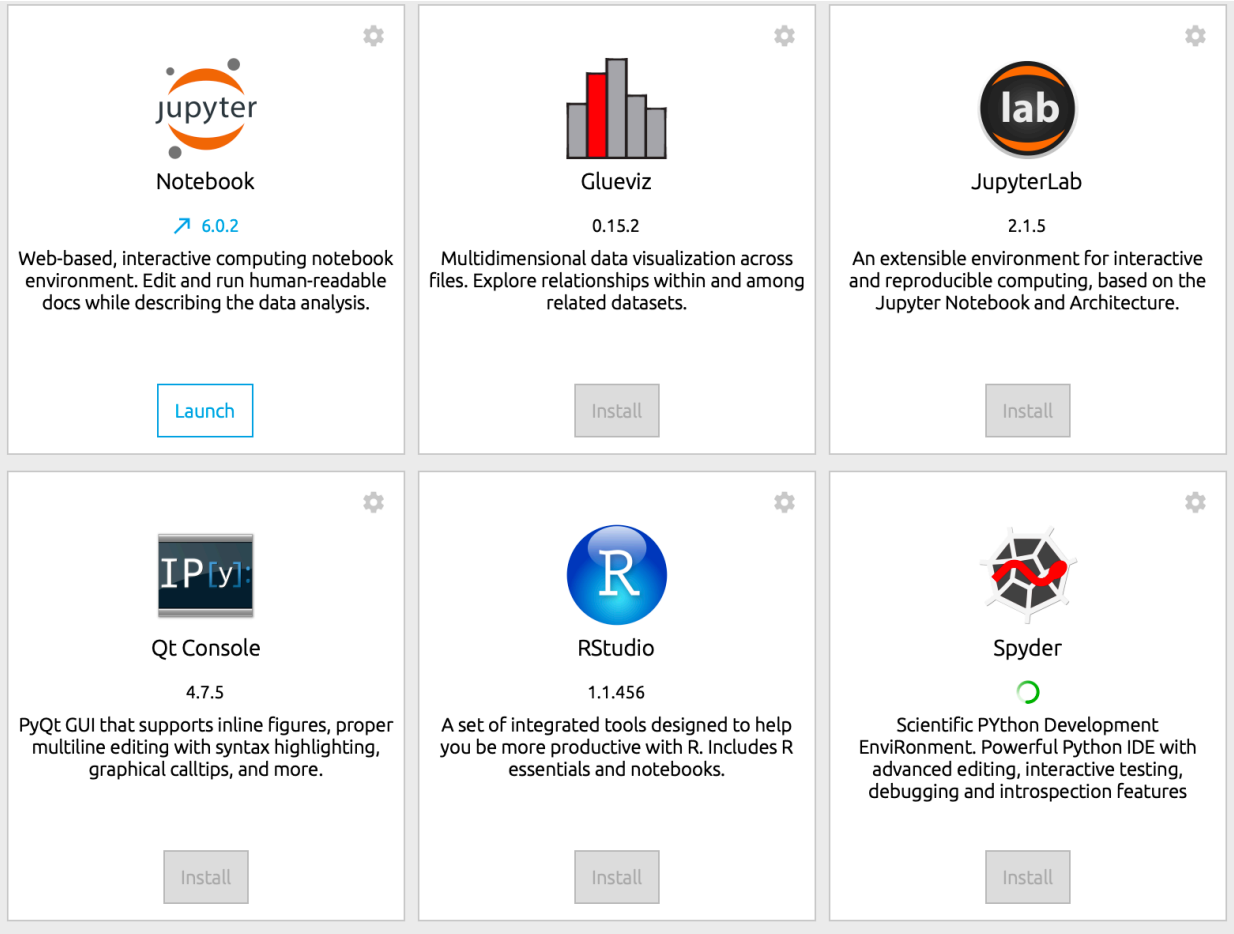
Search database nr using Blastp (protein-protein BLAST)

Show results in a new window

Percent identity 99.68% with Shingella

Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
MULTISPECIES: ribose-phosphate diphosphokinase [Proteobacteria]	635	635	100%	0.0	100.00%	WP_001298109.1
Chain A, Ribose-phosphate pyrophosphokinase [Escherichia coli O157:H7]	635	635	100%	0.0	100.00%	6ASV_A
ribose-phosphate pyrophosphokinase [Escherichia coli]	634	634	100%	0.0	99.68%	EFN7311852.1
ribose-phosphate diphosphokinase [Escherichia coli]	634	634	100%	0.0	99.68%	WP_114213719.1
ribose-phosphate diphosphokinase [Escherichia coli]	634	634	100%	0.0	99.68%	WP_137468128.1
ribose-phosphate pyrophosphokinase [Escherichia coli]	634	634	100%	0.0	99.68%	EEX8603463.1
ribose-phosphate diphosphokinase [Escherichia coli]	634	634	100%	0.0	99.68%	WP_004030744.1
TPA: ribose-phosphate pyrophosphokinase [Escherichia coli]	634	634	100%	0.0	99.68%	HAI4560783.1
ribose-phosphate pyrophosphokinase [Escherichia coli]	634	634	100%	0.0	99.68%	EGB0871914.1
ribose-phosphate diphosphokinase [Escherichia coli]	634	634	100%	0.0	99.68%	WP_169751754.1
PrsA [Shigella boydii]	634	634	100%	0.0	99.68%	ABC98167.1
TPA: ribose-phosphate pyrophosphokinase [Escherichia coli]	634	634	100%	0.0	99.68%	HAI2652416.1
ribose-phosphate diphosphokinase [Escherichia coli]	634	634	100%	0.0	99.68%	WP_058055490.1
ribose-phosphate diphosphokinase [Escherichia coli]	634	634	100%	0.0	99.68%	WP_044809305.1
ribose-phosphate pyrophosphokinase [Escherichia coli]	634	634	100%	0.0	99.68%	EFN9525760.1
TPA: ribose-phosphate pyrophosphokinase [Escherichia coli]	634	634	100%	0.0	99.68%	HAM7129415.1
MULTISPECIES: ribose-phosphate diphosphokinase [Shigella]	634	634	100%	0.0	99.68%	WP_012421410.1
ribose-phosphate pyrophosphokinase [Shigella sonnei]	634	634	100%	0.0	99.68%	EFZ4846886.1

Get python on your computer!



If you do not have python on your computer, check with me or your mentor!

Even if you have Anaconda, you may have to install Spyder:
But it messed up my Jupyter Notebooks!

To write a code outside of jupyter use:

A) Software for an Integrated Development Environment, like Spyder

B) A text editor designed for code editing, like emacs or vim

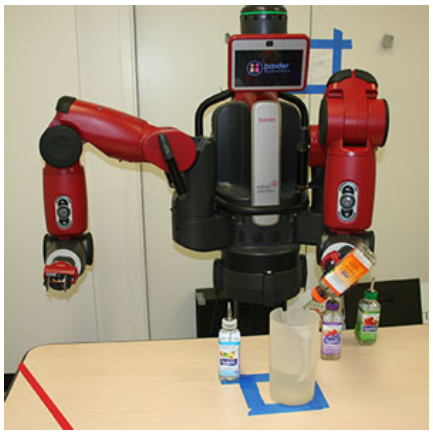
```
(base) adelstein@COSE-NICOLEAL-LT:~/Box/Teaching/306/HW/HW6$ vim aminoAcids.py
```

Pseudocode to practice while loops

- Imagine you have a robot who can go shopping for you!
  - I love cereal and milk, but I'm lactose intolerant.
  - My robot needs to find the lactose free, preferably organic, milk for me. If not found, get soy milk!
- Discuss with your group:
  - Why is `while` better than `for` loop?
  - Write some pseudo-code for the robot to get my milk or another item that your group chooses.

What syntax did you use?

A) if, elif, else
B) while & for loops
C) a variable
D) a function



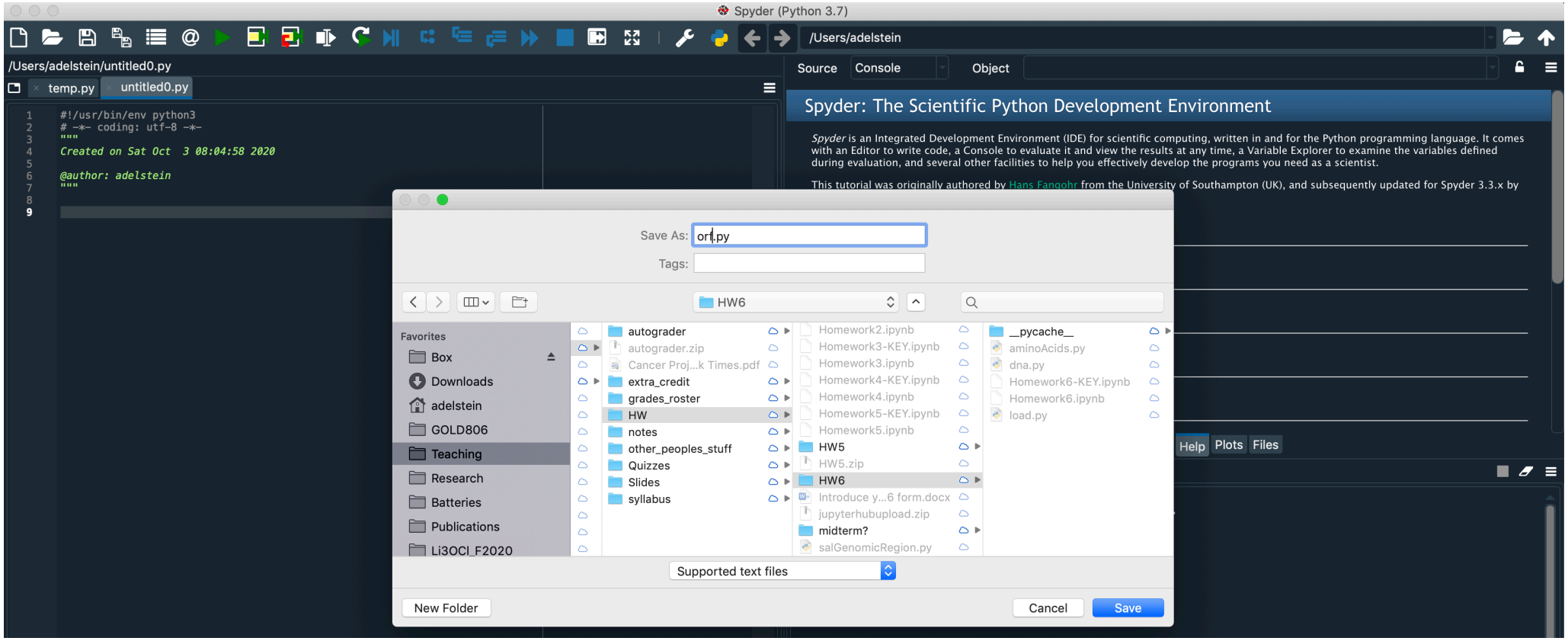
How to import your functions!

"All" of your code from HW 5 should be in a file called orf.py in order to import it!

```
def restOfORF(DNA):
```

```
def oneFrame(DNA):
```

```
def longestORF(DNA):
```



`orf.py` can call other codes!

Test functions in Jupyter OR Spyder.

For example, copy the showOff function.