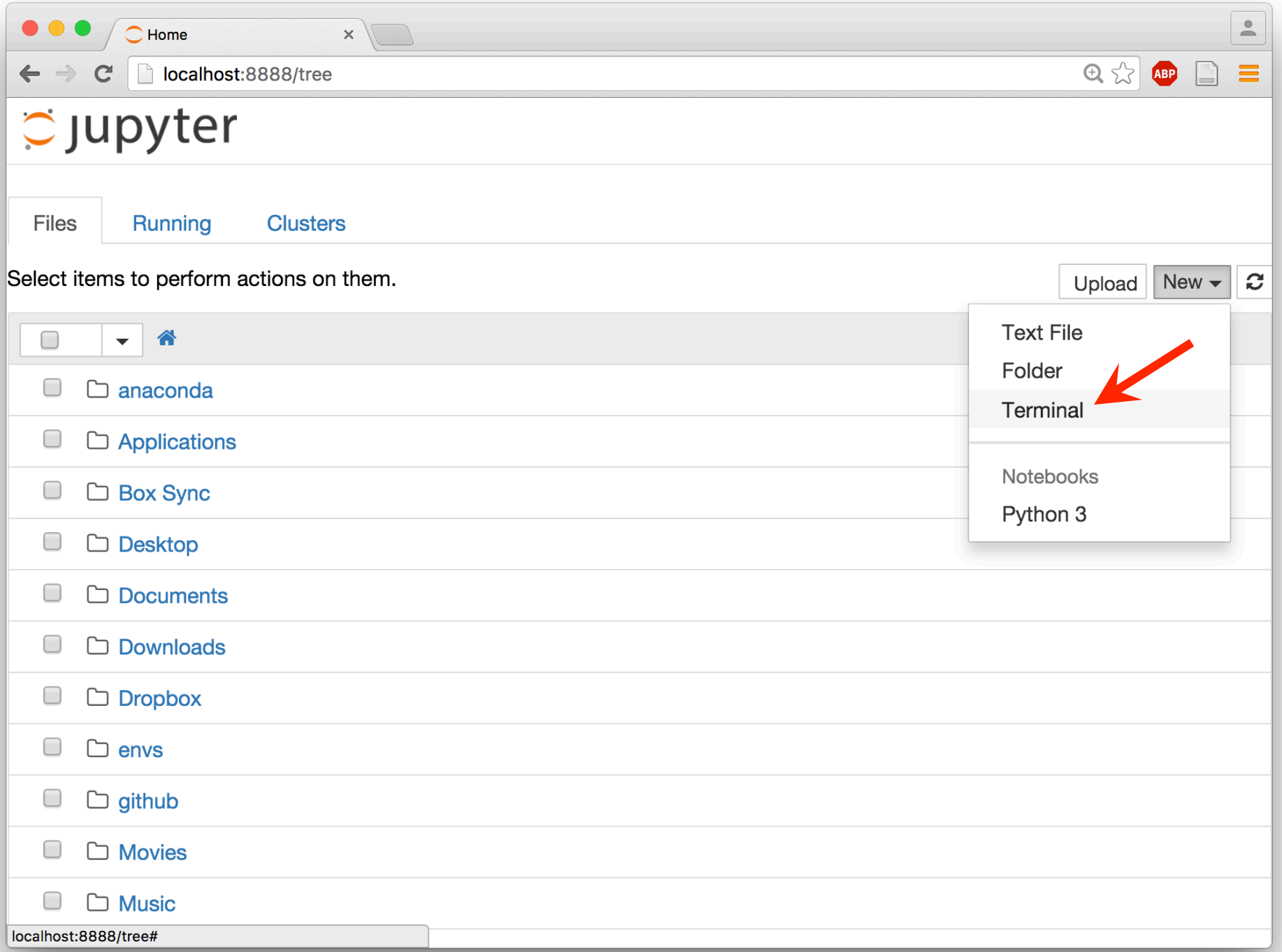
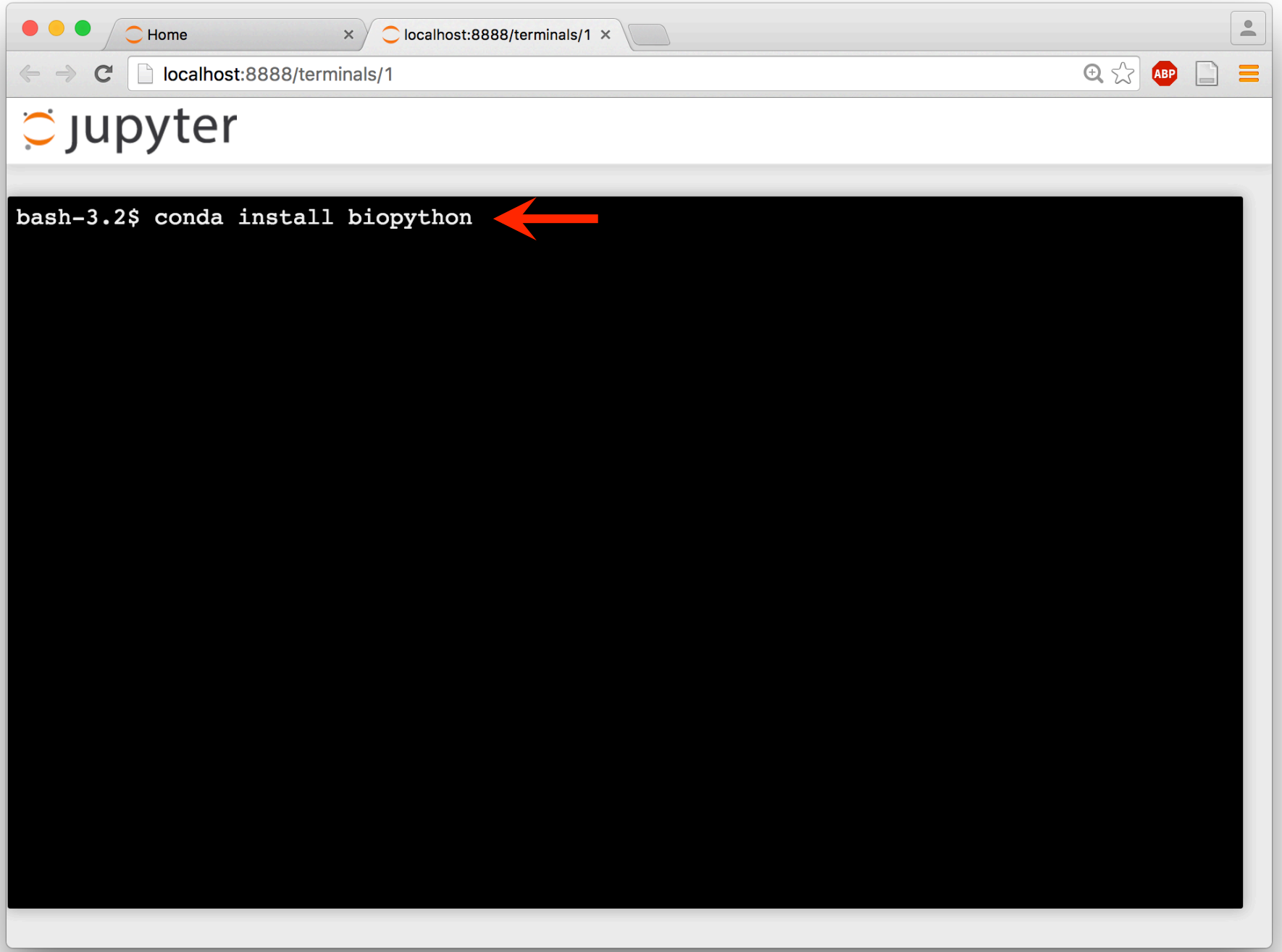
Working with biological sequence data

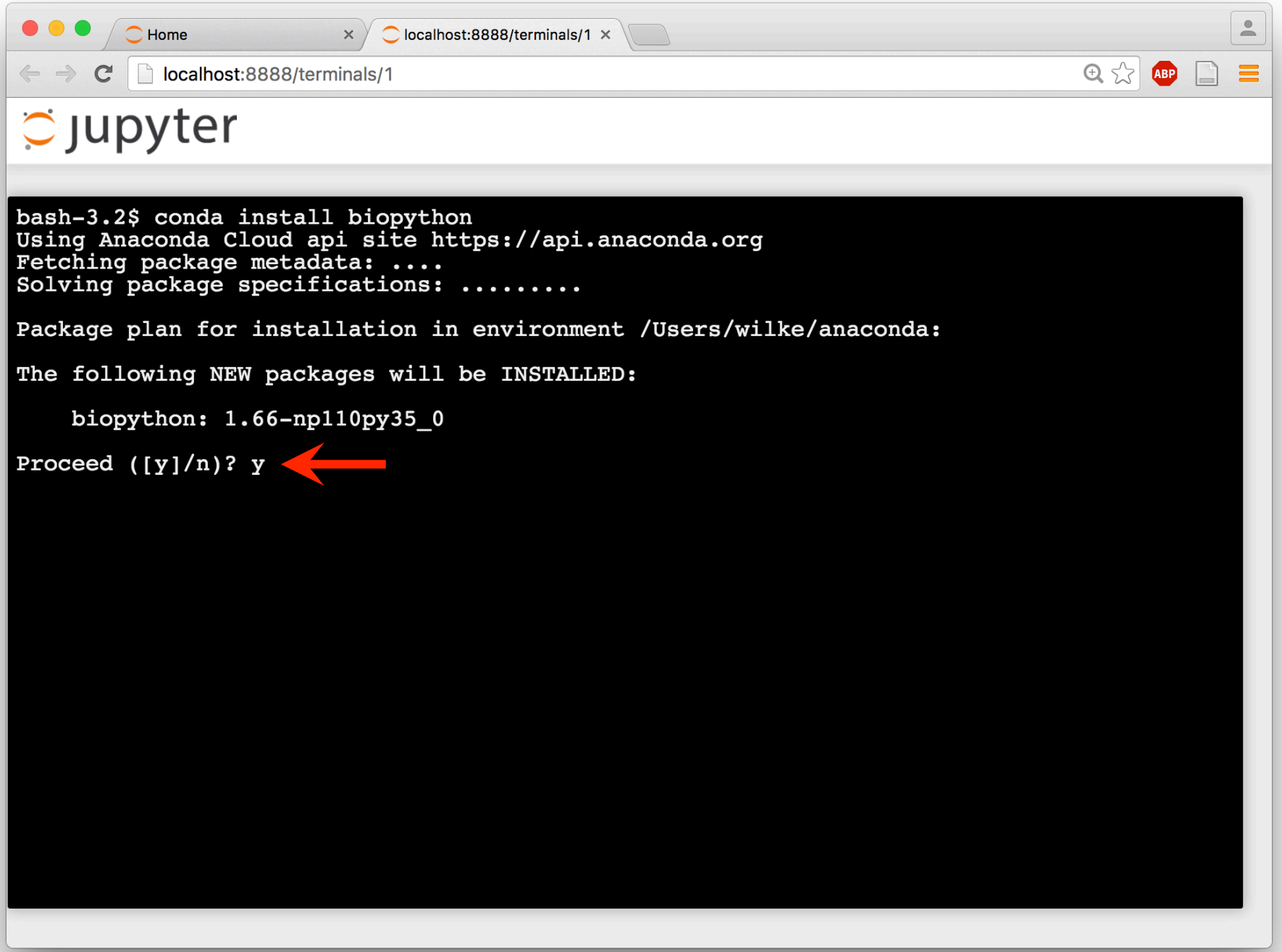
Installing Biopython:
We will need a command prompt/terminal

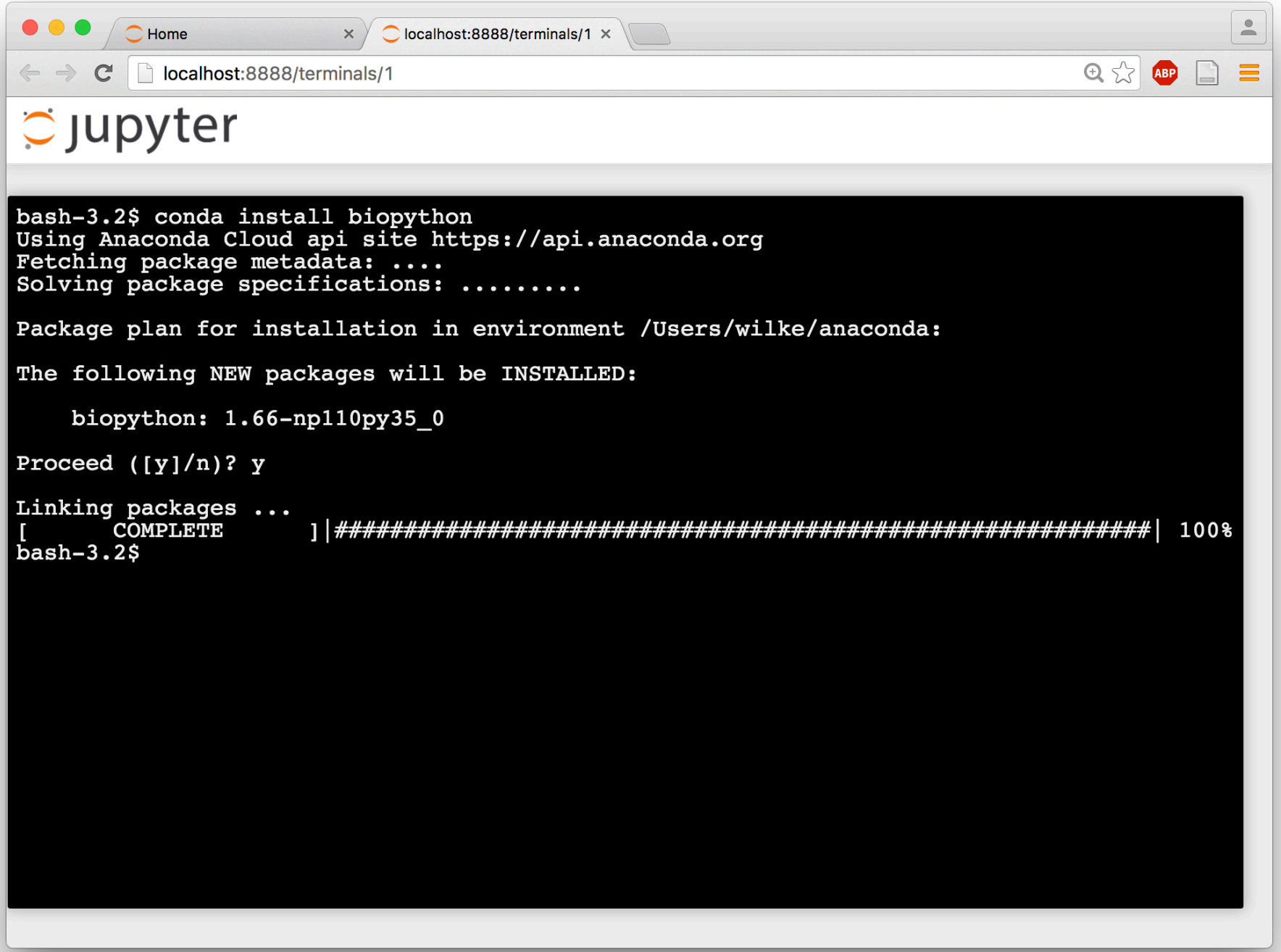
- On OS X:
  - Open terminal window
    or
  - Open terminal in Jupyter

- On Windows:
  - Go to "Start"
    click "Run"
    click "Command Prompt"


Home
localhost:8888/tree
jupyter
Files
Running
Clusters
Select items to perform actions on them.
Upload
New
Text File
Folder
Terminal
Notebooks
Python 3
anaconda
Applications
Box Sync
Desktop
Documents
Downloads
Dropbox
envs
github
Movies
Music
localhost:8888/tree#

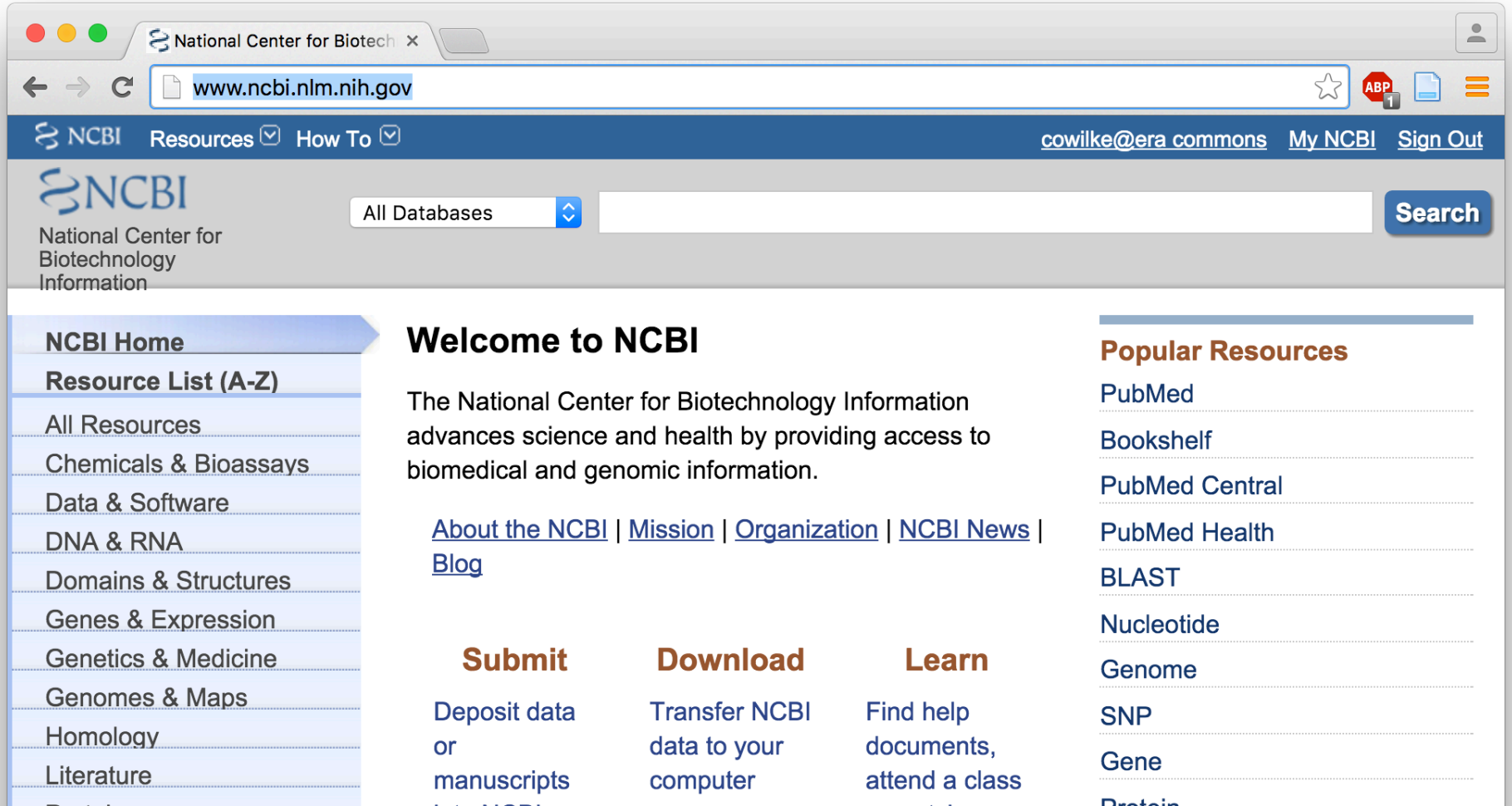

Home
localhost:8888/terminals/1
localhost:8888/terminals/1
jupyter
bash-3.2$ conda install biopython


Home
localhost:8888/terminals/1
localhost:8888/terminals/1
jupyter
bash-3.2$ conda install biopython
Using Anaconda Cloud api site https://api.anaconda.org
Fetching package metadata: ....
Solving package specifications: ....
Package plan for installation in environment /Users/wilke/anaconda:
The following NEW packages will be INSTALLED:
biopython: 1.66-np110py35_0
Proceed ([y]/n)? y

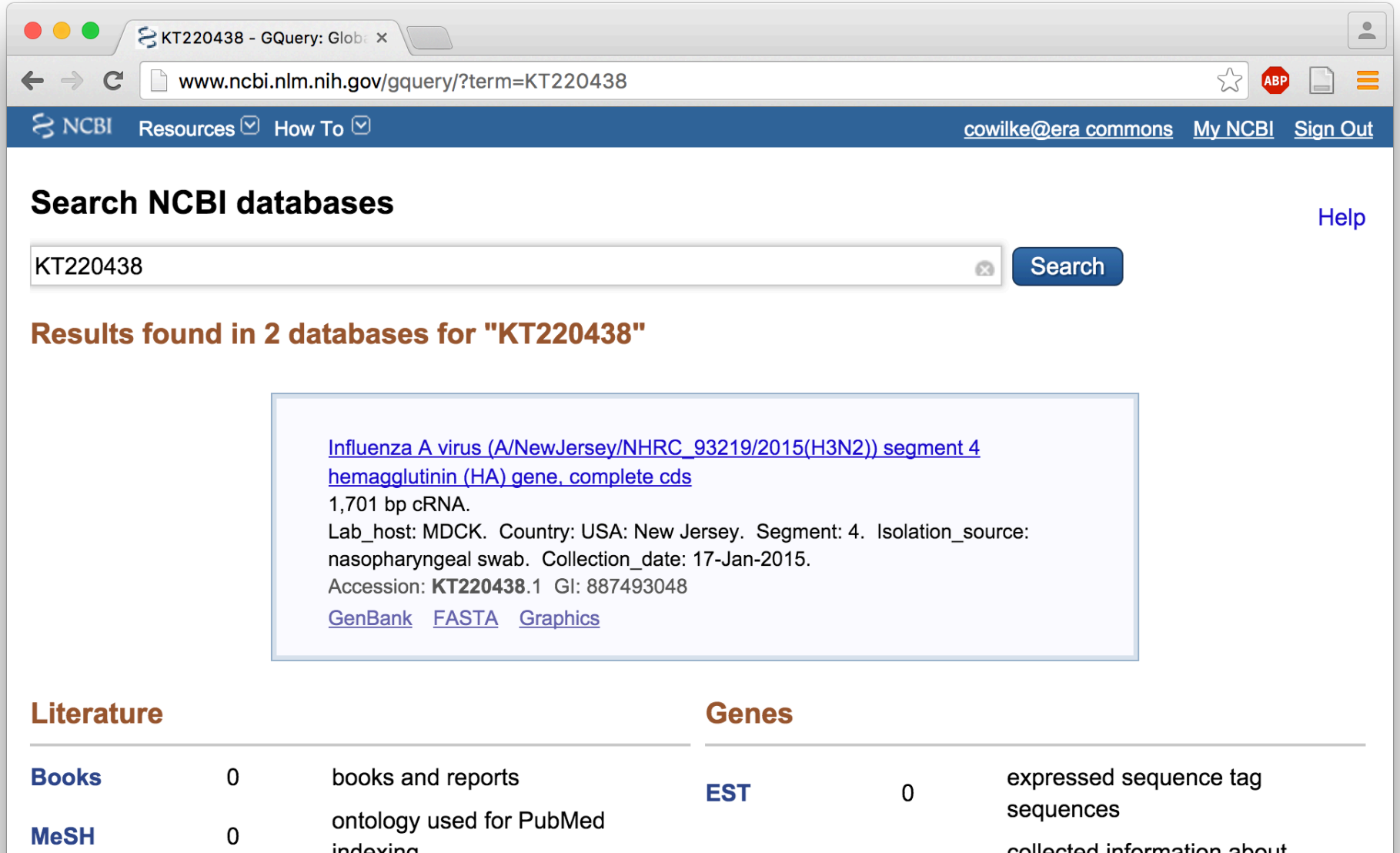

Home
localhost:8888/terminals/1
localhost:8888/terminals/1
jupyter
bash-3.2$ conda install biopython
Using Anaconda Cloud api site https://api.anaconda.org
Fetching package metadata: ....
Solving package specifications: ....
Package plan for installation in environment /Users/wilke/anaconda:
The following NEW packages will be INSTALLED:
biopython: 1.66-np110py35_0
Proceed ([y]/n)? y
Linking packages ...
[COMPLETE]|##########| 100%
bash-3.2$

Getting biological data: The NCBI databases

http://www.ncbi.nlm.nih.gov/



Try search for "KT220438"



Direct link to search results

http://www.ncbi.nlm.nih.gov/gquery/?term=KT220438

A genbank record is just a simple text file

```
LOCUS       KT220438                1701 bp    cRNA    linear   VRL 20-JUL-2015
DEFINITION  Influenza A virus (A/NewJersey/NHRC_93219/2015(H3N2)) segment 4
            hemagglutinin (HA) gene, complete cds.
ACCESSION   KT220438
VERSION     KT220438.1  GI:887493048
KEYWORDS    .
SOURCE      Influenza A virus (A/New Jersey/NHRC_93219/2015(H3N2))
  ORGANISM  Influenza A virus (A/New Jersey/NHRC_93219/2015(H3N2))
            Viruses; ssRNA viruses; ssRNA negative-strand viruses;
            Orthomyxoviridae; Influenzavirus A.
REFERENCE   1  (bases 1 to 1701)
  AUTHORS   Sitz,C.R., Thammavong,H.L., Balansay-Ames,M.S., Hawksworth,A.W.,
            Myers,C.A. and Brice,G.T.
  TITLE     GEISS Influenza Surveillance Response Program
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 1701)
  AUTHORS   Sitz,C.R., Thammavong,H.L., Balansay-Ames,M.S., Hawksworth,A.W.,
            Myers,C.A. and Brice,G.T.
  TITLE     Direct Submission
  JOURNAL   Submitted (29-JUN-2015) Operational Infectious Diseases, Naval
            Health Research Center, 140 Sylvester Rd., San Diego, CA 92106, USA
COMMENT     ##Assembly-Data-START##
            Sequencing Technology :: Sanger dideoxy sequencing
            ##Assembly-Data-END##
FEATURES             Location/Qualifiers
     source          1..1701
                     /organism="Influenza A virus (A/New
                     Jersey/NHRC_93219/2015(H3N2))"
                     /mol_type="viral cRNA"
                     /strain="A/NewJersey/NHRC_93219/2015"
                     /serotype="H3N2"
```

```
FEATURES             Location/Qualifiers
     source          1..1701
                     /organism="Influenza A virus (A/New
                     Jersey/NHRC_93219/2015(H3N2))"
                     /mol_type="viral cRNA"
                     /strain="A/NewJersey/NHRC_93219/2015"
                     /serotype="H3N2"
                     /isolation_source="nasopharyngeal swab"
                     /host="Homo sapiens"
                     /db_xref="taxon:1682360"
                     /segment="4"
                     /lab_host="MDCK"
                     /country="USA: New Jersey"
                     /collection_date="17-Jan-2015"
     gene            1..1701
                     /gene="HA"
     CDS             1..1701
                     /gene="HA"
                     /function="receptor binding and fusion protein"
                     /codon_start=1
                     /product="hemagglutinin"
                     /protein_id="AKQ43545.1"
                     /db_xref="GI:887493049"
                     /translation="MKTIIALSYILCLVFAQKIPGNDNSTATLCLGHHAVPNGTIVKT
                     ITNDRIEVTNATELVQNSSIGEICDSPHQILDGENCTLIDALLGDPQCDGFQNKKWDL
                     FVERSKAYSNCYPYDVPDYASLRSLVASSGTLEFNNESFNWTGVTQNGTSSACIRRSS
                     SSFFSRLNWLTHLNYTYPALNVTMPNNEQFDKLYIWGVHHPGTDKDQIFLYAQSSGRI
                     TVSTKRSQQAVIPNIGSRPRIRDIPSRISIYWTIVKPGDILLINSTGNLIAPRGYFKI
                     RSGKSSIMRSDAPIGKCKSECITPNGSIPNDKPFQNVNRITYGACPRYVKHSTLKLAT
                     GMRNVPEKQTRGIFGAIAGFIENGWEGMVDGWYGFRHQNSEGRGQAADLKSTQAAIDQ
                     INGKLNRLIGKTNEKFHQIEKEFSEVEGRIQDLEKYVEDTKIDLWSYNAELLVALENQ
```

ORIGIN

```
        1 atgaagacta tcattgcttt gagctacatt ctatgtctgg ttttcgctca aaaaattcct
       61 ggaaatgaca atagcacggc aacgctgtgc cttgggcacc atgcagtacc aaacggaacg
      121 atagtgaaaa caatcacaaa tgaccgaatt gaagttacta atgctactga gctggttcag
      181 aattcctcaa taggtgaaat atgcgacagt cctcatcaga tccttgatgg agaaaactgc
      241 acactaatag atgctctatt gggagaccct cagtgtgatg gctttcaaaa taagaaatgg
      301 gacctttttg ttgaacgaag caaagcctac agcaactgct acccttatga tgtgccggat
      361 tatgcctccc ttaggtcact agttgcctca tccggcacac tggagtttaa caatgaaagc
      421 ttcaattgga ctggagtcac tcaaaacgga acaagttctg cttgcataag gagatctagt
      481 agtagtttct ttagtagatt aaattggttg acccacttaa actacacata cccagcattg
      541 aacgtgacta tgccaaacaa tgaacaattt gacaaattgt acatttgggg ggttcaccac
      601 ccgggtacgg acaaggacca aatcttcctg tatgctcaat catcaggaag aatcacagta
      661 tctaccaaaa gaagccaaca agctgtaatc ccaaatatcg gatctagacc cagaataagg
      721 gatatcccta gcagaataag catctattgg acaatagtaa aaccgggaga catacttttg
      781 attaacagca cagggaatct aattgctcct aggggttact tcaaaatacg aagtgggaaa
      841 agctcaataa tgagatcaga tgcacccatt ggcaaatgca agtctgaatg catcactcca
      901 aatggaagca ttcccaatga caaaccattc caaaatgtaa acaggatcac atacggggcc
      961 tgtcccagat atgttaagca tagcactcta aaattggcaa caggaatgcg aaatgtacca
     1021 gagaaacaaa ctagaggcat atttggcgca atagcgggtt tcatagaaaa tggttgggag
     1081 ggaatggtgg atggttggta cggtttcagg catcaaaatt ctgagggaag aggacaagca
     1141 gcagatctca aaagcactca agcagcaatc gatcaaatca atgggaagct gaatcgattg
     1201 atcgggaaaa ccaacgagaa attccatcag attgaaaaag aattctcaga agtagaagga
     1261 agaattcagg accttgagaa atatgttgag gacactaaaa tagatctctg gtcatacaac
     1321 gcggagcttc ttgttgccct ggagaaccaa catacarttg atctaactga ctcagaaatg
     1381 aacaaactgt ttgaaaaaac aaagaagcaa ctgagggaaa atgctgagga tatgggaaat
     1441 ggttgtttca aaatatacca caaatgtgac aatgcctgca taggatcaat aagaaatgga
     1501 acttatgacc acaatgtgta cagggatgaa gcattaaaca accggttcca gatcaaggga
     1561 gttgagctga agtcagggta caaagattgg atcctatgga tttcctytgc catatcatgt
     1621 tttttgcttt gtgttgcttt gttggggttc atcatgtggg cctgccaaaa gggcaacatt
     1681 aggtgcaaca tttgcatttg a
```

//