In Class Protein Research Exercises 12.741

(Version November 2022)

These exercises are designed to provide some familiarity and practice with using omics datasets, to assist with the in class research project on ocean microbial proteins. The focus begins with the Ocean Portal Portal (OPP; www.oceanproteinportal.org) to discover proteins that are environmentally relevant (found in the environment) then progresses to incorporate using other related tools. As with all omic data, the possibilities are seemingly limitless, but this worksheet is intended to provide some practice and familiarity with working and interpreting the big data type and provide new biogeochemical understanding.

In class exercise: work your way through the following exercises. But since Alphafold can take 30-60 minutes, start with #5, and grab a sequence using #1 to grab an input sequence. Then let it run in the background while working your way through the other exercises.

1. **Using the Search bar and Protein Results table.** Using the www.oceanproteinportal.org site, search terms of "Protein Name" searches annotations of the proteins within the OPP and is the most effective search tool currently. Within the "Search Value" box try a few different iterations. The filter widgets on the left side can be left at their defaults (concentration, depth range, filter size, dataset, date range). Try a bunch of different search entries to see the kinds of results possible. Note that you can grab the sequence for other steps by clicking the "View Sequence" link at the right side of the protein table and highlighting and copying the text (make sure to expand the window and grab the full text).
    a. Try entering a enzyme name or fragment of an enzyme name (Rubisco, oxidoreductase, protease, etc.) You can browse and pick one from a the Enzyme Commission number database too such as from the Expasy EC catalog also here for the parent website. Alternatively you can browse BRENDA the comprehensive enzyme information system.
    b. Try entering a protein function, structure, sensor, or regulatory system (transporter, phage, kinase, regulator)
    c. Try entering an element (iron, nickel, zinc, sulfur etc)
    d. Try browsing the unknowns, or the "dark protein" often called "Hypothetical protein of unknown function" by searching "Hypothetical". If you figure out what one of these do it would be very helpful to everyone and be big news (full disclosure this might take you a few years to do it with biochemical and/or genetic tools).
2. **Interpreting the protein table results.**
    a. How many proteins are returned? Note the search is limited to 200 entries. Use a more specific search if >200. Note that you can select the number of entries shown in the dropdown box up to 200.
    b. Which proteins are most abundant? Look at the Spectral count column. The higher the number the more abundant the protein is. Clicking on the box reverses the order. Spectral counts are the counts of peptide identifications within a

dataset. ***Note that similar proteins can share counts.*** In other words, homologous proteins with very similar sequences (typically from the same species) that share some peptides but have slight differences in sequence (and peptides components) will have similar values as identical peptides are present within those similar sequences.

c. What organisms are making the proteins? Look in the NCBI Taxon column. Are there multiple species (e.g. *Prochlorococcus*)? Multiple strains (MIT9301 vs MIT 9312)?
d. What are the types of proteins that have returned from the search? This is the Product Name and KEGG columns which provide the top annotation information from BLAST searches.
e. What does the distribution of these proteins in the oceans look like? There are three several visualizations you can try.
    i. **View Section** will provide visualizations from the datasets that have enough data. These are depth sections across ocean transects (ProteOMZ and MetZyme currently) and provide a heatmap view of spectral count density. These are one of the most useful visualizations in the OPP currently that allow exploration of the depth and geographic distribution of proteins across thousands of kilometers of ocean.
    ii. **Profile Plot** provides a depth profile view of proteins. Note this can visualize multiple proteins through the check boxes. This tool is useful to explore the vertical distributions of similarly annotated proteins. Try clicking several with similar names and plotting them. Note that there is a hover-over and zoom in feature in this as well.
    iii. **Circle Map View** provides a from space top down view of relative protein abundance. Note this can visualize multiple proteins through the check boxes. The circle size is normalized to the total spectral counts at each depth, and different depths are overlaid. As a result the most abundant proteins will be most obvious in this visualization. Note that this feature is only present in the Beta Version 2 https:///kg.oceanproteinportal.org currently. This view is similar to the Ocean Gene Atlas view, and is useful to explore the geographic distribution across biogeochemical provinces. Future versions will

3. Run a **BLAST search** (sequence alignment) against NCBI GenBank. Starting at the Ocean Protein Portal grab a sequence of a protein of interest: Click View Sequence at the far right side of the Protein Table. A pop-up window will open with the corresponding protein sequence. You may need to enlarge the window to see the full sequence and the View NCBI Protein BLAST button. Click “View NCBI Protein Blast” - this will open a new tab for a BLASTp search at NCBI (also here https://blast.ncbi.nlm.nih.gov/Blast.cgi or just google NCBI Blast in the future) and prepopulate the sequence box. Scroll down and click the BLAST button. This search will take a few minutes. Once complete click through the tab headers over the results. Also ask Mak about his story about sequences to GEOTRACES.

a. On the “Descriptions” tab, note the protein Description (protein name/annotation), its taxonomic origin (Scientific Name), and some various scores, including the E-value (expected value; the smaller/lower the value the better the result), and percent identities (% identical peptides) and the length of the sequence in amino acids being compared to (Accession Length Acc. Len.).
b. The Graphic Summary tab provides a color coded view of the coverage of your “query” sequence against the sequence pulled up by the BLAST search. If most of the sequence is covered and has a high score that indicates many similar sequences in the database.
c. The Alignments tab lets you see the pairwise alignments (query versus subject, numbered by amino acid location) with the letters between the query and subject indicating shared or similar sequence. in paircan also explore the taxonomy and alignments on this page. NCBI has a lot of embedded information, so feel free to explore.
d. Download sequences: Maybe you want to grab the subject sequence for further queries - it is likely from a cultivated organism rather than from the natural environment and hence is better understood or could be used for experiments. From the Descriptions tab you can select individual proteins and then the Download tab for the FASTA sequence text files.
e. Distance Trees of Results. Phylogenetic trees are tools used to interpret the evolutionary relatedness, or homology, of sequences. You can create a tree of these BLASTp results by clicking the “Distance tree of results” link (both in “other reports” line in the header and in the table. Note that this is using the 200 queries that the BLASTp output was originally limited to. What are you looking at? Note the small scale bar in the bottom right corner, the longer the horizontal branch the larger amount of change in the sequences. You can click on the nodes to collapse them, or to examine the sequence alignment of them. Note that creating trees is often considered a bit of an art form that involves trimming sequences and choosing the appropriate phylogenetic model. But this tool allows you to make some quick assessments about relationships between the sequences you are exploring.

4. **Ocean Gene Atlas**: This is a database and portal for metagenomic and metatranscriptomic results primarily from the Tara expedition. Copy a sequence from the Ocean Protein Portal (View sequence, and highlight the text in the box and use control-C to copy and control-V to paste (or apple equivalent). Then navigate to the Ocean Gene Atlas (OGS) https://tara-oceans.mio.osupytheas.fr/ocean-gene-atlas/ and paste the sequence into the “Either, query sequence:” box. The OGS defaults to using its prokaryotic metagenomic findings in the Dataset box, use this unless you know you are targeting a Eukaryotic protein. You can leave the other settings as default and hit Submit. It may take a few minutes for the search to run. The results show circle plots of gene abundance at different depths (SRF surface, DCM deep chlorophyll maximum, MES mesopelagic, MIX mixed layer). Not that the dataset is primarily euphotic zone samples. Multiple size fractions can be explored as well. Clicking on the circles provides a taxonomic breakdown of the sample. Comparisons to a few environmental parameters

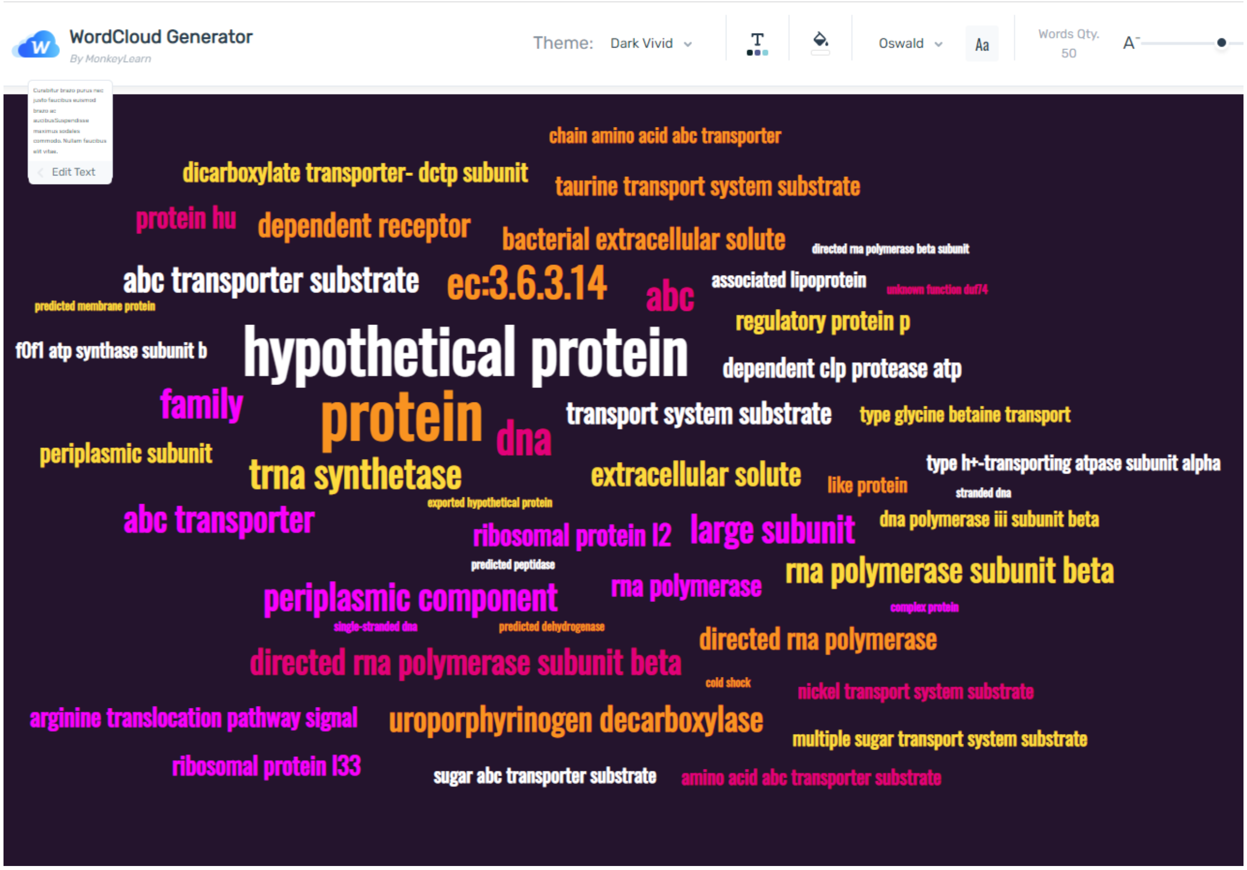
can be conducted in the next set of panels below. Now you can say you have explored metagenomic data too.

5. **AlphaFold** Predict protein structure using Alphafold machine learning. Computer based protein structure determination has improved greatly with the application of machine learning. You can employ Alphafold using a google Colab (python) notebook and a simplified version of Alphafold. https://colab.research.google.com/github/deepmind/alphafold/blob/main/notebooks/AlphaFold.ipynb You will need access to a google account to access Colab. The process is simple after that. Click the following link paste in your sequence into the third cell. The select the Runtime drop down menu then select Run All. The program will likely take 30 minutes or longer to complete. This version works for monomers (single proteins). Running multimers (multi-protein complex) is more computationally expensive and requires running locally or buying additional compute power from google. Paste in a sequence from the Ocean Protein Portal View Sequence option. Examine the structure, not that you can spin them and zoom in and out. Also examine where they produce regions of high and low confidence. Do you believe the structure? Take a Screenshot of it and try another protein sequence from the OPP. Note that the results are also automatically downloaded to your computer in a zip file.
6. PDB (**Protein Data Bank**) is a portal for structural information about proteins, and includes experimentally determined structures using x-ray crystallography, and now computationally predicted structures. Navibate to https://www.rcsb.org/ Copy the sequence from the OPP website and paste it into the search query. Examine the results that are returned and assess whether it is expected. Explore the 3D view of the protein. You can click on the 3D view and explore the structure of the protein. Searches can also be done using protein names, try that as well (carbonic anhydrase; superoxide dismutases also see word cloud below). There are often many options, including mutants where a single amino acid has been changed to see the effect on the structure and binding sites. Sometimes metal sites are included in structures, but often not, and often a different metal is inserted experimentally.
7. **EXPASY** (https://www.expasy.org/) is a bioinformatics portal with many useful tools. Perform some simple sequence manipulations like getting formation about molecular weight (MW) or cleaving your proteins. To get information about MW, go to the “proteins and proteomes” tab on the lefthand side of the page. Then, select the “Compute PI/MW” box option and then “Browse this resource”. Plug in your sequence from the OPP from step 3 or 4 above to get the theoretical PI and PW. You can also use the “PeptideCutter” to cleave proteins into smaller peptides by enzyme digestion. To do this, select the “PeptideCutter” and then “browse the resource”. Paste in your sequence here as well. Notice how many enzymes can cut proteins, typically at specific locations. Trypsin is the enzyme almost exclusively in proteomics because it cleaves at K or R residues which produce mass spectrometry amenable peptides. Select trypsin and run the search. Examine the results of peptides that have been cleaved. What sites were they cleaved at? Redo the cleaving using “all available enzymes”. What do the numbers after the enzymes refer too? Reflect on how the carbon cycle depends on proteases like these

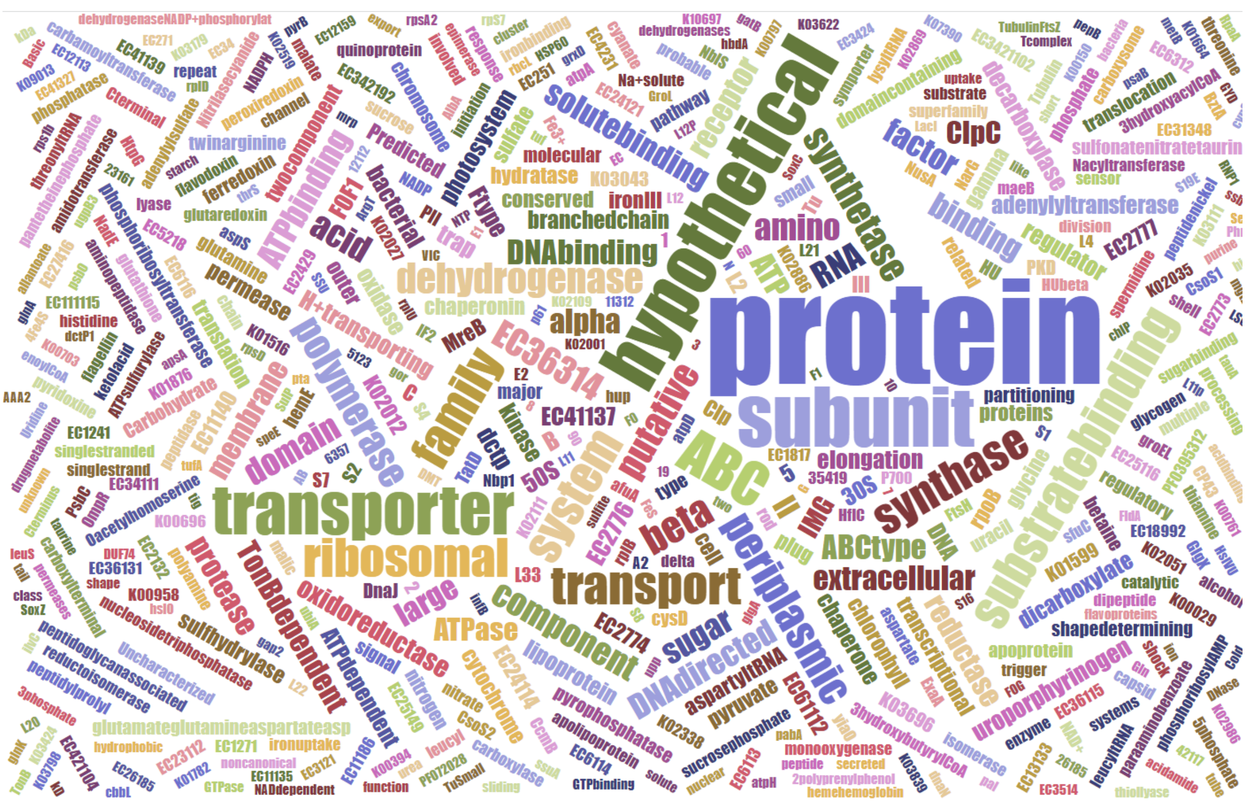
(minus the N and C terminal proteases that aren't included here) to cleave proteins into amino acids as the first step of N remineralization.

Congratulations! You have explored multi-omic meta-omics datasets. Oo-la-la! How many tabs do you have open? That's normal. This is a rapidly evolving space with new tools and websites being modified continuously, but you now have some basic experience finding sequences of interest and using them to explore various portals and repositories to understand the distirbution of proteins and genes in nature.

Breaking word cloud generators with omic data. Use for keyword search inspiration:



WordCloud Generator
By MonkeyLearn
Theme: Dark Vivid
Oswald
Aa
Words Qty. 50
Edit Text
chain amino acid abc transporter
dicarboxylate transporter- dctp subunit
taurine transport system substrate
protein hu
dependent receptor
bacterial extracellular solute
directed rna polymerase beta subunit
abc transporter substrate
ec:3.6.3.14
abc
associated lipoprotein
unknown function duf74
predicted membrane protein
regulatory protein p
f0f1 atp synthase subunit b
hypothetical protein
dependent clp protease atp
family
protein
dna
transport system substrate
type glycine betaine transport
periplasmic subunit
trna synthetase
extracellular solute
like protein
type h+-transporting atpase subunit alpha
stranded dna
exported hypothetical protein
abc transporter
ribosomal protein l2
large subunit
dna polymerase iii subunit beta
predicted peptidase
periplasmic component
rna polymerase
rna polymerase subunit beta
complex protein
single-stranded dna
predicted dehydrogenase
directed rna polymerase
directed rna polymerase subunit beta
cold shock
nickel transport system substrate
arginine translocation pathway signal
uroporphyrinogen decarboxylase
multiple sugar transport system substrate
ribosomal protein l33
sugar abc transporter substrate
amino acid abc transporter substrate



protein
hypothetical
subunit
transporter
ribosomal
transport
synthase
substratebinding
periplasmic
synthetase
family
system
putative
component
polymerase
dehydrogenase
DNAbinding
ABC
ABCtype
extracellular
TonBdependent
Oxidoreductase
binding
factor
ClpC
Solutebinding
receptor
decarboxylase
uroporphyrinogen
DNAdirected
membrane
domain
acid
beta
alpha
large
ATPase
sugar
lipoprotein
chaperonin
amino
RNA
elongation
partitioning
proteins
adenylyltransferase
sulfonatenitratetaurine
Nacyltransferase
translocation
regulator
branchedchain
conserved
molecular
hydratase
ironIII
twinarginine
ferredoxin
flavodoxin
glutaredoxin
permease
glutamine
membrane
protease
chaperone
chlorophyll
transcriptional
reductase
apoprotein
trigger
shapedetermining
dicarboxylate
dipeptide
catalytic
glycogen
regulatory
betaine
quinoprotein
repeat
phosphatase
Cterminal
Carbamoyltransferase
aminotransferase
phosphoribosyltransferase
histidine
singlestranded
singlestrand
Carbohydrate
Oacetylhomoserine
shape
class
peptidylprolyl
reductoisomerase
nucleosidetriphosphatase
sulfhydrylase
Uncharacterized
ATPdependent
glutamateglutamineaspartateasp
ironuptake
hydrophobic
noncanonical
NADdependent
function
cytochrome
nitrate
pyrophosphatase
apolipoprotein
sucrosephosphate
aspartyltRNA
pyruvate
monooxygenase
peptide
secreted
2polyprenylphenol
hemehemoglobin
Isomerase
enzyme
systems
paraaminobenzoate
acidamide
thiolyase
GTPbinding
Nbp1
50S
30S
EC36314
EC41137
EC2774
EC18992
EC31348
K03043
K02001
MreB
TatD
DnaJ
cysD
delta
type
photosystem
Predicted
bacterial
sulfate
pathway
Na+solute
chromosome
dehydrogenaseNADP+phosphorylat
dehydrogenases
TubulinFtsZ
Tcomplex
substrate
uptake
superfamily
sensor
division
PKD
HUbeta