Discovering Microbial Protein Function in the Oceans
In class research Marine Bioinorganic Chemistry
12.741 2022 Document version 10/13/22

Overview of Research Motivation:
The oceans are a major contributor to the Earth's life support system, and the marine microbes that live within it are the living catalysts that promote Earth's life-sustaining biogeochemical cycles. In recent years scientists have begun to accumulate a tremendous amount of data on the proteins synthesized by marine life and publicly shared through the Ocean Protein Portal. Currently there are over 100,000 proteins in the Ocean Protein Portal, most of which have not been examined in terms of their ecological and biogeochemical roles.

In this project we will ask the basic science questions: What are the interesting and important components of this system across the ocean regions and depths? What is the function, distribution, and taxonomic source of various proteins within the ocean? Can their distribution be explained through oceanographic, biochemical, and taxonomic contexts? Can the role of uncultivated organisms be better understood through the study of their deployed proteins? More detailed questions can also be asked for example: What is the diversity of a specific type of protein (enzyme or transporter for example) and how does its distribution change across biogeochemical and physical provinces? This in-class research project aims to fulfill the following educational and research objectives, providing the students with a real-life big data research experience where there is no right or wrong answer.

Educational Goals:
1. Overall goal: Provide a hands-on in silico research experience where students forge connections between biochemistry and biogeochemistry
2. Specific activities and goals:
    a. Gain experience and comfort exploring large omic datasets using OPP search results:
        i. Launch text and ID based search queries (KEGG, PFam, E.C.)
        ii. Understand and interpret "Protein Found" table output
        iii. Create visualizations of protein distributions (section, profile, and circle maps)
    b. Connect proteins datasets to marine microbe taxonomic origin
        i. Conduct bioinformatic analyses of primary sequence and its relationship to other organisms' sequences using BLAST
        ii. Construct and interpret phylogenetic tree in NCBI
    c. Connect knowledge of protein function(s) (as described in classroom lectures) with research examples:
        i. Assess function based on protein annotation and literature search
        ii. Make critical assessment of quality of computer based annotations
        iii. Connect observations to any relevant data within BCO-DMO repositories, exploring environmental datasets
        iv. Use PFam, KEGG, Brenda,or Uniprot databases to infer function

v. Compare patterns of corresponding gene distribution within metagenomic resources (Ocean Gene Atlas or IMG)
vi. Hypothesize subcellular localization based on Expasy output data (cytoplasmic versus membrane localization)

d. Obtain (if available) or build (if tractable) a model protein 3D structure using PDB, PyMol, or Alphafold, predict the metal binding site based on primary and tertiary sequence. This may not be available/tractable for large protein complexes.
e. Interpret protein distribution within oceanographic context
    i. Conduct literature search about the connection between a protein and biogeochemical cycle(s)
    ii. Compare distributions to oceanographic biogeochemical provinces, nutrient fields, temperature gradients etc.
f. Improve and practice concise and articulate technical writing skills (Schimel guide)

Approach (map onto in-class and out of class sessions):
Sites: www.oceanproteinportal.org (Elasticsearch production site)
And beta V2 site (not public): KG Protein Portal

1. Have each student identify and characterize a protein component in the oceans based on discussion of protein functions
2. Determine the protein's location in the ocean
3. Examine the homology and phylogenetic relationships of protein(s)
4. Research could focus on proteins as organized by Enzyme Commission Groups, regulatory systems, or transporters.
5. Write a 3 paragraphs for website publication with figures from the above activities that synthesizes information from the biochemical level to the biogeochemical level

Timelines Goals:
Lecture: overview of proteins and their roles in biology and biogeochemistry (2-3 lectures).
In classroom: hands-on efforts:
Introduction to Ocean Protein Portal and capabilities (i-iv above)
In discussion with Instructor and TA: students will select protein of interest, and begin to explore outside of class time.
End products: By end of semester present results orally and prepare written blog information for website publication.

Protein Selection:
One of the challenges of this project is choosing a protein of interest to characterize. What classifies as interesting? This is subjective and there is no wrong answer. It will take a little exploration and critical analysis to determine if the protein is correctly annotated. Also some quality control screening will be needed to determine if there is sufficient data to make an interesting story. Finally some creative thinking and comparison with environmental data will be needed to place protein data into an oceanographic context. Below is a list of potential example

proteins for study. Keyword searches work best and rely on the computer generated annotations of the submitted datasets. Element or nutrient names work well as keyword searches.

Example proteins for exploration:

Biomarkers for nutrient stress:
Nitrogen: Urea transporter, NtcA, P-II, ammonia transporter
Phosphorus: Alkaline phosphatase, PstS (phosphate transporter)
Iron: Flavodoxin/Ferredoxin, Plastocyanin, iron transporters
Zinc: ZCRP-A, ZCRP-B
Vitamin B12: CBA1, MetH/MetE
See Walworth et al., 2022 for biomarker review.

Enzymes:
Ammonia monooxygenase
Nitrite oxidoreductase
Nitrogenase
Rubisco
Catalase
Alcohol dehydrogenase
Protease/Peptidase
Blue copper proteins
Many many others - explore KEGG, PFam, E.C. or organism papers for additional ideas
See Saunders et al., 2022 for overview of KEGG and E.C. groups found

Transporters:
TonB dependent transporters (many)
ABC transporters (many)
Specific molecular transporters (ammonia, phosphate, divalent cation,
*transporters can be difficult for computers to annotate, but can be specific and show compelling distributions.
Review on marine TonB by Hopkinson and Barbeau, 2012. Review on gut TonB.

Sensors/Regulators:
Phosphate (PhoB)
Nitrogen (P-II, NtcA)
Fur (iron)
Zur (Zn)
Review on marine two component regulators by Held et al., 2019

Structural:
Phage capsid proteins (cyanophage and other)
Carboxysomes

Storage:
Ferritin
Bacterioferritin
Metallothionein
Cyanophycin (N)
Glycogen (C)

Unknowns (for the brave!):
Hypothetical protein of unknown function

Ocean Protein Research Grading Rubric

Name:_____

Ocean Protein Portal Project

Criteria	Excellent (9-10)	Good (7-8)	Satisfactory (5-6)	Needs improvement (0-4)
Baseline Content	The content is accurate and includes a discussion of the following: who makes the protein, what its function is, and how it is distributed in the ocean. There is a thorough discussion linking the function and who makes the protein to its global distribution	The content is accurate and includes a discussion of the following: who makes the protein, what its function is, and how it is distributed in the ocean. There is a brief discussion linking the function and who makes the protein to its global distribution	The content is missing one of the following: who makes the protein, what its function is, and how it is distributed in the ocean. The report may contain minor misinformation. There is a brief discussion linking the function and who makes the protein to its global distribution	The content is missing two or more of the following: who makes the protein, what its function is, and how it is distributed in the ocean. The report may contain misinformation. There is no discussion linking function and who makes the protein to global distribution is lacking
Visualization	Accurate heatmaps, section plot and/or circle maps, and a depth profile are included in the report. The figures are well-labeled and referred to in the written portion of the text. Interpretations of	One of the required figures (heatmaps, section plot and/or circle maps, and a depth profile) is missing without an adequate reason for not including the figure. The figures are well labeled and referred to in the written portion of	One of the required figures (heatmaps, section plot and/or circle maps, and a depth profile) is missing without an adequate reason for not including the figure. The figures are not well labeled but are referred to in the text. Interpretations	Two or more of the required figures (heatmaps, section plot and/or circle maps, and a depth profile) are not included. The figures are not well labeled or not linked to the written text. There are few, if

	these figures are linked back to the baseline content.	the text. Interpretations of these figures are linked back to the baseline content.	of this figure is linked back to the baseline content.	any, interpretations.
Use of external sources	The report incorporates several external sources (e.g. PDB, NCDI, Kegg, Wikipedia, literature, etc). The content from the external source is well-integrated with the information from the OPP. All external references are correctly cited.	The report incorporates several external sources (e.g. PDB, NCDI, Kegg, Wikipedia, literature, etc). The content from the external source is adequately integrated with information from OPP. All external references are correctly cited.	The report incorporates two external sources (e.g. PDB, NCDI, Kegg, Wikipedia, literature, etc). The content from the external source is somewhat integrated with information from OPP. External references are incorrectly cited.	The report incorporates only one external source (e.g. PDB, NCDI, Kegg, Wikipedia, literature, etc). The content from the external source is not adequately integrated with information from OPP. External references are incorrectly cited.
Discussion of protein structure	A figure of the protein structure is given. Additionally, in the text, there is a detailed discussion about the structure of the protein using alpha fold or pyMOL. Details may include where the metal atoms are located, coordination chemistry, description of ligands, and written depiction of 3D structure.	A figure of the protein structure is given. There is a limited discussion about the structure of the protein using alpha fold or pyMOL. Discussion includes description of the 3D structure. *If there is no structural information for your molecule, using a similar molecule as a proxy is sufficient.	The 3D structure is included with little discussion of the structure. *If there is no structural information for your molecule, using a similar molecule as a proxy is sufficient.	There is no discussion of 3D protein structures.

	*If there is no structural information for your molecule, using a similar molecule as a proxy is sufficient.			
Written clarity	The sections and figures are well-integrated. The text is generally accessible to the general scientific community (i.e. non-biologists and non-oceanographers). Any jargon that is used is well defined. There are very few, if any, grammatical errors.	The sections and figures are well-integrated and easy to follow. Some of the jargon is not well-defined. There are few grammatical errors.	The sections and figures are tied together but jump around logically. Some jargon is used without explanation. There are some grammatical errors.	Sections and figures are not well-integrated. It is hard for follow the text and the report is inaccessible to a non-expert. Jargon is used without definition. There are many grammatical errors.

Total: _____/50 points =