ENGS 107: Problem Set #5: Project Outline and Video Pitch

TASK:

Produce a script and an up to ten minutes video (based on the script) to pitch your project idea.

Problem Definition

To mitigate the rising threats posed to human and environmental health by global warming, the International Energy Agency has developed the Net Zero Emissions by 2050 Scenario [1-4]. This framework outlines a path for the global energy sector to achieve net zero carbon emissions and limit global temperature rise to 1.5°C by 2050 [2,4]. The transportation sector accounts for 23% of the total greenhouse-gas emissions from the global energy sector [5]. Heavy-duty transportation (which includes airplanes, trucks, trains, and ships) are a primary source of these emissions [6-7].

In an effort to decarbonize the transportation sector, the IEA has projected that the production of liquid biofuels must increase about 2.6-fold by 2030 [8]. Biologically derived cellulosic ethanol exhibits potential as a liquid fuel to meet the rising demand for liquid biofuels, as ethanol can be catalytically upgraded to advanced fuel molecules for use in diesel and jet biofuels to displace fossil fuels [9-11]. **However, achieving industrially relevant ethanol titers is a foundational problem limiting the commercialization of cellulosic biofuel production.**

Question

Microbes that can natively break down cellulosic biomass, such as corn stover, are a logical choice to engineer for cellulosic ethanol production [12-13]. One prominent example is *Clostridium thermocellum*, a thermophilic anaerobic microbe that grows optimally at high temperatures and in the absence of oxygen [12,14]. To engineer *C. thermocellum* for high titer cellulosic ethanol production, we need to understand the factors that limit ethanol titer in this microbe. However, to date we do not fully understand these factors. **This raises the fundamental question: What factors are limiting high ethanol titer in *Clostridium thermocellum*?**

State-of-the-art

Current state-of-the-art approaches aimed at improving ethanol titer in *C. thermocellum* are drawn from the field of metabolic engineering, which is aimed at modifying the metabolism of a microbe through genetic and molecular biology approaches to improve its ability to produce a biochemical of interest. One state-of-the-art approach aimed at improving ethanol titer in *C. thermocellum* is adaptive evolution. This involves repeated culturing of a microbe under conditions that favor robust growth, which enables the microbe to acquire mutations that improve its product formation ability. In a recent study, two rounds of adaptive evolution were performed with *C. thermocellum*, but the best producing strain only achieved an ethanol titer of 22.4 g/L, which is far below the benchmark ethanol titer of 40 g/L required for commercialization [15-16].

Additional state-of-the-art approaches aimed at increasing ethanol titer in *C. thermocellum* have included: expressing enzymes from other microbes in *C. thermocellum* to improve chemical reactions in the ethanol production pathway and deleting branching metabolic pathways to divert more carbon flux to ethanol [17-19]. These approaches aimed at increasing

ethanol titer in *C. thermocellum* have failed to develop a strain capable of producing 40 g/L ethanol [17-19]. This indicates a demand for an alternative approach to understand the factors that limit high ethanol titer in *C. thermocellum*.

New Approach

One approach is to identify a thermophilic pyruvate transporter. This is a protein embedded in the cell membrane that enables the transport of pyruvate into the cell under high temperature conditions. Pyruvate is a compound of high interest, because it functions as a branching point in the central metabolism of *C. thermocellum* en route to ethanol from cellulosic biomass [20-21]. Expressing a thermophilic pyruvate transporter in *C. thermocellum* would enable the study of pyruvate transport into the cell, which may inform whether the metabolic bottlenecks limiting high ethanol titer are either upstream or downstream of pyruvate. Elucidating the relative location of these metabolic bottlenecks would be a key step forward in understanding the factors that limit high ethanol titer in *C. thermocellum*. However, no thermophilic pyruvate transporters are known to exist.

Nexus that Enables the Project

The absence of a known thermophilic pyruvate transporter provides an opportunity and a **nexus** to apply Bayesian modeling as a computational approach to identify a thermophilic pyruvate transporter to subsequently experimentally test for activity in *C. thermocellum*.

Hypothesis

For this project, I will use Bayesian inference to computationally test the following **hypothesis**: Is this protein a thermophilic pyruvate transporter?

Methods

Which methods

To identify a thermophilic pyruvate transporter, my **methodological approach** is broadly focused on using a Bayesian statistical model with classification to computationally produce a posterior probability for each examined protein [22-27]. The key steps of this method are outlined here:

The **first step** is to perform data collection and feature extraction. To do this, I will first create a positive dataset of mesophilic (or room temperature) pyruvate transporters, a negative dataset of non-pyruvate transporters in thermophilic microbes, and a candidate dataset of thermophilic membrane proteins [24]. For feature extraction, I will focus on extracting features that are indicative of potential functionality as a thermophilic pyruvate transporter including: indicators of thermophilicity, structure-based features, sequence-based features, and genomic context features [26].

The **second step** will be to define and build my Bayesian model based on my collected data and extracted features from step 1. To do this, I will first define the prior distribution, which represents the initial belief about whether a protein is a thermophilic pyruvate transporter. The prior will be based on available knowledge including: the probability that a transporter moves pyruvate, the probability of a protein being a thermophilic transporter, and functional and

structural properties that are common to transport proteins. Second, I will define the likelihood function, which will describe how probable it is to observe the extracted features if a protein is truly a thermophilic pyruvate transporter. I plan to model the likelihood function using a Gaussian process, as this enables uncertainty quantification. Third, I will define the posterior probability using Bayes' theorem. This posterior probability will quantify the belief that a protein is a thermophilic pyruvate transporter after incorporating new data [25].

The **third step** will be to use a Markov Chain Monte Carlo method to generate samples and obtain a posterior probability distribution for each examined protein [25].

The **fourth step** will be to produce a ranked list of candidate proteins based on their posterior probability and confidence levels of being a thermophilic pyruvate transporter.

The **fifth step** will be to select the top 10-15 candidates from the ranked list and experimentally test these proteins for activity in *C. thermocellum*.

If any successful protein candidates are found to enable pyruvate transporter, the **final step of my method** will be to update my prior with this data to improve my model accuracy.

Why this choice

I selected this computational methodology for several reasons. First, Bayesian inference enables the integration of prior knowledge (such as data about mesophilic pyruvate transporters) to inform my model. Second, instead of simply generating a binary classification, a Bayesian model produces posterior probabilities, which enable the ranking of candidate proteins based on confidence levels. Third, since pyruvate transport is a complex biological phenomenon, this approach enables the incorporation of diverse biological datasets to improve model accuracy. Fourth, using a Markov Chain Monte Carlo method facilitates an accurate posterior probability estimation for each examined protein, which subsequently assists in the ranking of candidate proteins. And finally, this approach allows for experimental results to be integrated back into the Bayesian model as priors to improve its accuracy [25].

What is your data?

As eluded to earlier, **my data** for this analysis will come from several different protein, sequence, structure, thermophilicity, and annotation databases. To generate my positive, negative, and candidate datasets I will acquire my data on transport proteins from the KEGG database, the SwissProt section of the UniProt database, and the NCBI genome database. For feature extraction, I will extract protein sequence features from the NCBI BLAST database, structural features from AlphaFold and Swiss-Model databases, thermophilicity data from the ThermoProt Database, and genomic context features from the KEGG pathways database.

Expected Results

In terms of **expected results**, it is expected that this analysis will produce a ranked list of 10-15 candidate proteins that are worth experimentally testing for functionality as a thermophilic pyruvate transporter in *C. thermocellum*. Of the tested proteins, it is expected that 2-3 will exhibit functionality as a thermophilic pyruvate transporter.

Intellectual Merit

In terms of **intellectual merit**, this project exhibits potential to advance our understanding of the factors that limit ethanol titer in *C. thermocellum* in several regards. First, this research may lead to the identification of a thermophilic pyruvate transporter. This may inform what specific structural motifs or amino acid sequences enable functionality under high temperatures. This knowledge could be applied to inform metabolic engineering efforts aimed at engineering transporters for use in thermophilic microbes.

Second, this work may provide insights into where the carbon from pyruvate is routed through metabolism in different strains of *C. thermocellum*, advancing our understanding of its non-canonical metabolism [21].

Third, expressing a thermophilic pyruvate transporter in *C. thermocellum* may inform whether the metabolic bottlenecks limiting ethanol titer are located upstream or downstream of pyruvate. This may inform future research aimed at resolving these bottlenecks, leading to higher ethanol titers in *C. thermocellum*.

Fourth, this work provides a framework for using Bayesian modeling to computationally identify proteins for diverse applications in thermophilic microbes.

Broader Impacts

In terms of **broader impacts,** this work may provide key steps forward towards developing a *C. thermocellum* strain capable of producing cellulosic ethanol at commercial scales. This furthers our efforts aimed at developing an industrial scale biorefinery for cellulosic ethanol production to displace fossil fuels from the transportation sector and achieve net zero emissions by 2050 [4,12].

Video Duration: 9 minutes 59 seconds (1 second less than 10 minute requirement)

*Please note that both the .mp4 video file and PDF script file are located on my GitHub Repository for ENGS 107. The link to access my GitHub Repository is provided below: https://github.com/isaiah-richardson28/Dartmouth-ENGS-107-Winter-2025

Also please note that no code script was uploaded for this problem set as no code script was 1) required per discussion with Dr. Keller in class nor 2) written to produce any of the figures, plots, or graphs used in this video. All figures, plots, and graphs were produced on Microsoft Excel, Adobe Illustrator, Microsoft PowerPoint, or were acquired from the Internet and cited.

References (Citations)

1. World Health Organization. *Climate Change*. Retrieved from https://www.who.int/news-room/fact-sheets/detail/climate-change-and-health, (2023).

2. International Energy Agency. *Global Energy and Climate Model*. Retrieved from https://www.iea.org/reports/global-energy-and-climate- model/net-zero-emissions-by-2050-scenario-nz, License: CC BY 4.0 (2023).

3. Keller, K., Helgeson, C. & Srikrishnan, V. Climate Risk Management. *Annual Review of Earth and Planetary Sciences* **49**, 95-116 (2021).

4. Intergovernmental Panel on Climate Change. *Climate Change 2023 Synthesis Report: Summary for Policymakers*. Retrieved from https://www.ipcc.ch/report/ar6/syr/ (2023).

5. IEA. *Global energy-related CO2 emissions by sector*. Retrieved from https://www.iea.org/data-and-statistics/charts/global-energy-related-co2-emissions-by-sector (2020).

6. IEA. *Net Zero by 2050: A Roadmap for the Global Energy Sector*. Retrieved from https://www.iea.org/reports/net-zero-by-2050 (2021).

7. IEA. *Global CO2 emissions from transport by sub-sector in the Net Zero Scenario, 2000-2030*. Retrieved from https://www.iea.org/data-and-statistics/charts/global-co2-emissions-from-transport-by-sub-sector-in-the-net-zero-scenario-2000-2030-2 (2023).

8. IEA. *Bioenergy use by sector globally in the Net Zero Scenario, 2010-2030*. Retrieved from https://www.iea.org/data-and-statistics/charts/bioenergy-use-by-sector-globally-in-the-net-zero-scenario-2010-2030 (2024).

9. Kubis, M.R. & Lynd, L.R. Carbon capture from corn stover ethanol production via maturing consolidating bioprocessing enables large negative biorefinery GHG emissions and fossil fuel-competitive economics. *Sustainable Energy & Fuels* **7**, 3842-3852 (2023).

10. Lynd, L.R., Beckham, G.T., Guss, A.M., Jayakody, L.N., Karp, E.M., Maranas, C., McCormick, R.L. et al. Toward lost-cost biological and hybrid biological/catalytic conversion of cellulosic biomass to fuels. *Energy & Environmental Science* **15**, 938-990 (2022).

11. Eagan, N.M., Kumbhalkar, M.D., Buchanan, S.J., Dumesic, J.A. & Huber, G.W. Chemistries and processes for the conversion of ethanol into middle-distillate fuels. *Nature Reviews* **3**, 223-249 (2019).

12. Lynd, L.R., Liang, X., Biddy, M.J., Allee, A., Cai, H., Foust, T., Himmel, M.E., Laser, M.S., Wang, M. & Wyman, C.E. Cellulosic Ethanol: Status and Innovation. *Current Opinion in Biotechnology* **45**, 202-211 (2017).

13. Zheng, B., Yu, S., Chen, Z. & Huo, Y.X. A consolidated review of commercial-scale high-value products from lignocellulosic biomass. *Frontiers in Microbiology* **13**, 1-21 (2022).

14. Akinosho, H., Yee, K., Close, D. & Ragauskas, A. The emergence of Clostridium thermocellum as a high utility candidate for consolidated bioprocessing applications. *Frontiers in Chemistry* **2**, 1-14 (2014).

15. Tian, L., Papanek, B., Olson, D.G., Rydzak, T., Holwerda, E.K., Zheng, T., Zhou, J. et al. Simultaneous achievement of high ethanol yield and titer in *Clostridium thermocellum*. *Biotechnology for Biofuels and Bioproducts* **9**, 1-11 (2016).

16. Dien, B.S., Cotta, M.A. & Jeffries, T.W. Bacteria engineering for fuel ethanol production: current status. *Applied Microbiology and Biotechnology* **63**, 258-266 (2003).

17. Tian, L., Perot, S.J., Hon, S., Zhou, J., Liang, X., Bouvier, J.T., Guss, A.M., Olson, D.G. & Lynd, L.R. Enhanced ethanol formation by *Clostridium thermocellum* via pyruvate decarboxylase. *Microbial Cell Factories* **16**, 1-10 (2017).

18. Hon, S., Holwerda, E.K., Worthen, R.S., Maloney, M.I., Tian, L., Cui, J., Lin, P.P. et al. Expressing the *Thermoanaerobacterium saccharolyticum pforA* in engineered *Clostridium thermocellum* improves ethanol production. *Biotechnology for Biofuels and Bioproducts* **11**, 1-11 (2018).

19. Holwerda, E.K., Olson, D.G., Ruppertsberger, N.M., Stevenson, D.M., Murphy, S.J.L., Maloney, M.I., Lanahan, A.A. et al. Metabolic and evolutionary responses of *Clostridium thermocellum* to genetic interventions aimed at improving ethanol production. *Biotechnology for Biofuels and Bioenergy* **13**, 1-20 (2020).

20. Olson, D.G., Hörl, M., Fuhrer, T., Cui, J., Zhou, J., Maloney, M.I., Amador-Noguez, D., Tian, L., Sauer, U. & Lynd, L.R. Glycolysis without pyruvate kinase in Clostridium thermocellum. *Metabolic Engineering* **39**, 1690180 (2017).

21. Zhou, J., Olson, D.G., Argyros, D.A., Deng, Y., van Gulik, W.M., van Dijken, J.P. & Lynd, L.R. Atypical Glycolysis in *Clostridium thermocellum*. *Applied and Environmental Microbiology* **88**, 1-15 (2022).

22. Alterovitz, G., Liu, J., Afkhami, E. & Ramoni, M.F. Bayesian methods for proteomics. *Proteomics* **7**, 2843-2855 (2007).

23. Wilkinson, D.J. Bayesian methods in bioinformatics and computational systems biology. *Briefings in Bioinformatics* **8**, 109-116 (2007).

24. Ijaq, J., Malik, G., Kumar, A., Das, P.S., Meena, N., Bethi, N., Sundararajan, V.S. & Suravajhala, P. A model to predict the function of hypothetical proteins through a nine-point classification scoring schema. *BMC Bioinformatics* **20**, 1-8 (2019).

25. Crook, O.M., Chung, C.W. & Deane, C.M. Challenges and opportunities for Bayesian Statistics in Proteomics. *Journal of Proteome* **21**, 849-864 (2022).

26. Wang, Y., Zhang, H., Zhong, H. & Xue, Z. Protein domain identification methods and online resources. *Computational and Structural Biotechnology* **19**, 1145-1153 (2021).

27. Li, Y.F., Arnold, R. J., Radivojac, P. & Tang, H. Protein identification problem from a Bayesian point of view. *Stat Interface* **5**, 21-37 (2012).