ENGS 107: Problem Set #6: Project Presentation Video

Problem Definition: Rising greenhouse gas emissions are accelerating climate change, posing risks to humans and the environment [1-4]. The transportation sector accounts for almost 1/4 of the greenhouse gases from the global energy sector [5]. To decarbonize transportation, the International Energy Agency has projected that liquid biofuel production has to increase 2.6-fold by 2030 [6]. Ethanol produced from cellulose (cellulosic ethanol) exhibits potential to meet the rising demand for liquid biofuels [7]. Cellulose is an abundant natural resource found in plants [8]. Ethanol can be efficiently converted into biofuels to displace fossil fuels [9-11]. However, achieving high ethanol concentrations (or titers) is a **problem** limiting the commercialization of cellulosic biofuels [12].

Question: Microbes such as *Clostridium thermocellum* that can naturally break down plant-based matter (or cellulosic biomass), are a logical choice to engineer for cellulosic ethanol production [12-13]. *C. thermocellum* grows at high temperatures and in the absence of oxygen, making it a thermophile and an anaerobe [13]. To date we do not fully understand the factors that limit high ethanol titer in *C. thermocellum* [12]. This raises the **question**: What factors are limiting high ethanol titer in this microbe?

State-of-the-art: State-of-the art approaches aimed at improving ethanol titer in *C. thermocellum* have included: performing experiments that enable mutations to arise that improve its ability to make ethanol, expressing enzymes (or catalytic proteins) from other microbes to improve inefficient chemical reactions, and deleting branching pathways to divert more carbon to ethanol [14-17]. These approaches have fallen short of developing a strain that can produce 40 g/L ethanol, the benchmark titer required for commercialization [14-18].

Solution: One potential **solution** to advance our understanding of the factors that limit ethanol titer in *C. thermocellum* is to identify a thermophilic pyruvate transporter (or TPT). This is a protein embedded in the membrane of the cell that enables the transport of pyruvate into the microbe under high temperatures. Pyruvate is a compound of interest, because it is a key intermediate in the metabolic pathway from cellobiose to ethanol in *C. thermocellum* [19-21]. Expressing a TPT in *C. thermocellum* would enable the study of pyruvate transport into the cell and its impact on ethanol production. This would help to identify if the factors limiting high ethanol titer are located upstream or downstream of pyruvate [20-21].

Nexus that Enables the Project:
To date, I am not aware of any TPTs that exist. Constructing a model based on Bayesian inference in R provides an opportunity and a **nexus** to identify a TPT by computationally generating a ranked list of thousands of candidate proteins based on predictions, experimentally testing the top candidates for activity in *C. thermocellum*, and iteratively refining the model structure based on experimental results.

Hypothesis: I **hypothesize** that candidate proteins predicted by the model to have high TPT activity will enable pyruvate uptake in *C. thermocellum*.

Methods (Which Methods): My method for identifying a TPT is divided into two Aims [22-27]. **Aim 1** is to generate an initial ranked list of candidates and experimentally test the top candidates for

activity in *C. thermocellum* (Keller, K., personal communication, February 2025). The **first step** is to define the model, where y is the yield (or prediction), X is a vector representing five different biological features, theta is the parameters, and epsilon is the model-observation mismatch. This step also includes defining the data generating function, the priors on the parameters, and the candidate function to calculate yield. The **second step** is to collect the feature data for each candidate from biological databases **followed by** calculating the feature prediction (or yield) for each candidate normalized between 0 and 1. The **fourth step** is to generate a ranked list of candidates based on yields, and **finally** experimentally test the top 30 candidates from the ranked list for activity in *C. thermocellum*.

A ranked list of 10 candidates generated by my model is shown in this table. The proteins are listed on the left, with the five feature values for each protein in the middle, and the calculated yield and rank on the right (Keller, K., personal communication, February 2025).

Aim 2 is to refine the model structure and ranking based on the experimental results (Keller, K., personal communication, February 2025). After obtaining the initial results, the **sixth step** is to define and calculate the likelihood function, **followed by** calibrating the model and refining the model structure and choices. This includes updating the priors on the parameters and running MCMC to generate a posterior distribution of parameters. The **next step** is to calculate the posterior predictions and re-rank the candidates. The **final step** is to select a new set of top candidates to test. This method can iterate through steps 5-9 to further refine the model. However, it is important to note that the effect of neglecting structural uncertainty is unknown because only a single model was used.

Why this choice: One reason for selecting this method is that it reduces experimental burden by computationally informing which candidates to test. Second, unlike a supervised machine learning approach, this Bayesian approach does not require known examples of TPTs as input data. Third, this Bayesian approach enables the incorporation of biological intuition as priors and refinement of the model structure as results are obtained. Fourth, this Bayesian approach learns which biological features are most predictive, which improves prediction accuracy.

Expected results: It is expected that 1-3 of the top 30 candidates from the initial ranked list will be able to transport pyruvate into *C. thermocellum*, while most will exhibit no activity. It is also expected that thermostability and sequence homology will be the most predictive features for determining TPT activity. And third, it is expected that prediction accuracy will improve in each iteration, and that 3-5 functional TPTs will be identified after a few iterations.

Intellectual merit: In terms of **intellectual merit**, this research may lead to the identification of a functional TPT. This may inform what specific features enable pyruvate transport under high temperatures. It may also inform where the carbon from pyruvate is routed through metabolism in *C. thermocellum* [20-21]. And it may also inform whether the factors limiting ethanol titer are located upstream or downstream of pyruvate [21]. This may inform future research aimed at resolving these problems and improving ethanol titer.

Broader impacts: In terms of **broader impacts,** this work provides a computational framework for identifying unknown transport proteins for diverse applications. It may also provide a key step

forward towards developing a strain of *C. thermocellum* that can produce cellulosic ethanol at an industrial scale. Finally, this work advances our efforts aimed at establishing a commercial biorefinery for cellulosic ethanol production to displace fossil fuels and achieve net zero emissions by 2050 [4,7].

Important Notes:

*Please note that this PDF script file is located on my ENGS 107 GitHub Repository and is named: [isaiah.d.richardson.th@dartmouth.edu].PS#6.pdf. The .MP4 video file was too large to add to the GitHub Repository (even when exported from PowerPoint in a low resolution format). Attached here is the link to access my GitHub Repository: https://github.com/isaiah-richardson28/Dartmouth-ENGS-107-Winter-2025

*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 office hours before class on 02/21/2025 nor 2) required per the Problem Set #6 description. The table of initial results shown on slide 9 of the video was prepared in Microsoft Excel and was adapted from a set of initial results that I obtained from a practice code script that I wrote in R for 10 hypothetical pyruvate transport proteins from mesophilic microbes. The final code script that I wrote in R with all of my initial results is located on my GitHub Repository and was submitted with Problem Set #7 (the proposal).

*Please note that references 22-27 were used to help initially guide the conceptual design of my method. However, the design of the method was primarily informed through personal communication with Dr. Keller during several in-person office hour meetings both before class (ECSC 042) and in his office (Irving 385) throughout February 2025 as cited in the script.

*Please note that references 28-30 are not cited in the video script but instead are cited within the slides of the video as they are references for images that were acquired from the Internet.

References

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

2. IEA. Global Energy and Climate Model. https://www.iea.org/reports/global-energy-and-climate-model. (2024).

3. Keller, K., Helgeson, C. & Srikrishnan, V. Climate Risk Management. *Annual Review of Earth and Planetary Sciences* **49**, 95-116. https://doi.org/10.1146/annurev-earth-080320-055847. (2021).

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

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

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

7. 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. https://doi.org/10.1016/j.copbio.2017.03.008. (2017).

8. 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. https://doi.org/10.3389/fmicb.2022.933882. (2022).

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. https://doi.org/10.1039/D3SE00353A. (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. https://doi.org/10.1039/D1EE02540F. (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. https://doi.org/10.1038/s41570-019-0084-4. (2019).

12. Olson, D.G., Maloney, M.I., Lanahan, A.A., Cervenka, N.D., Xia, Y., Pech-Canul, A., Hon, S., et al. Ethanol tolerance in engineered strains of *Clostridium thermocellum*. *Biotechnology for Biofuels and Bioproducts* **16**, 1-15. https://doi.org/10.1186/s13068-023-02379-z. (2023).

13. 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. https://doi.org/10.3389/fchem.2014.00066. (2014).

14. 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. https://doi.org/10.1186/s13068-016-0528-8. (2016).

15. 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. https://doi.org/10.1186/s12934-017-0783-9. (2017).

16. 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. https://doi.org/10.1186/s13068-018-1245-2. (2018).

17. 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. https://doi.org/10.1186/s13068-020-01680-5. (2020).

18. Dien, B.S., Cotta, M.A. & Jeffries, T.W. Bacteria engineering for fuel ethanol production: current status. *Applied Microbiology and Biotechnology* **63**, 258-266. https:///doi.org/10.1007/s00253-003-1444-y. (2003).

19. 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**, 169-180. https://doi.org/10.1016/j.ymben.2016.11.011. (2017).

20. 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* **79**, 3000-3008. https://doi.org/10.1128/AEM.04037-12. (2013).

21. Cui, J., Stevenson, D., Korosh, T., Amador-Noguez, D., Olson, D.G. & Lynd, L.R. Developing a Cell-Free Extract Reaction (CFER) System in *Clostridium thermocellum* to Identify Metabolic Limitations to Ethanol Production. *Frontiers in Energy Research* **8**, 1-15. https://doi.org/10.3389/fenrg.2020.00072. (2020).

22. Alterovitz, G., Liu, J., Afkhami, E. & Ramoni, M.F. Bayesian methods for proteomics. *Proteomics* **7**, 2843-2855. https://doi.org/10.1002/pmic.200700422. (2007).

23. Wilkinson, D.J. Bayesian methods in bioinformatics and computational systems biology. *Briefings in Bioinformatics* **8**, 109-116. https://doi.org/10.1093/bib/bbm007. (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. https://doi.org/10.1186/s12859-018-2554-y. (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. https://doi.org/10.1021/acs.jproteome.1c00859. (2022).

26. Wang, Y., Zhang, H., Zhong, H. & Xue, Z. Protein domain identification methods and online resources. *Computational and Structural Biotechnology* **19**, 1145-1153. https://doi.org/10.1016/j.csbj.2021.01.041. (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. https://dx.doi.org/10.4310/SII.2012.v5.n1.a3. (2012).

28. Freepik. "Green grass vector for cartoon." https://www.freepik.com/premium-vector/green-grass-vector-cartoon_290836915.htm. (No date).

29. Freepik. "Computer Cartoon Images." https://www.freepik.com/free-photos-vectors/computer-cartoon. (No date).

30. Freepik. "Biorefinery Images." https://www.freepik.com/free-photos-vectors/biorefinery. (No date).