Figure List for thioesterase ML manuscript:

12-12-20

Figure 1: Overview of type II fatty acid biosynthesis

Emphasize role of thioesterase in determining chain-length specificity

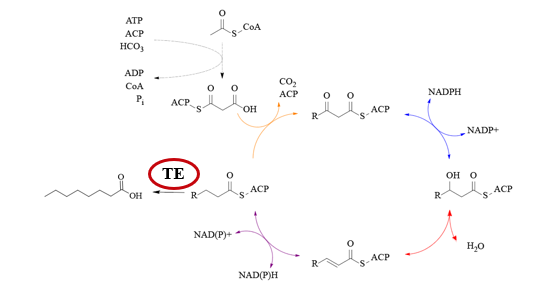


Figure 2A: Machine learning demonstration with training set

Not sure what is best to show here. Maybe a bar chart of accuracy improving with each iteration of the model.

Figure 2B: Graphical representation of SVM 🡪 flow chart of inputs and outputs of ensemble model maybe

Figure 3: Bioprospecting for a novel C10-specific thioesterase

-Select all uncharacterized thioesterases from *Cuphea* species with C10 acid in their seed oils (*lanceolata, llavea, procumbens, viscosissima*) 🡪 Run ML prediction of thioesterase specificity 🡪 synthesis and test 30 of the thioesterases in *E. coli* to 1) validate the model, and 2) bioprospect for the C10 thioesterase while we are at it

Potential Figure 3B: Bioprospecting with more traditional methods

-Since our hypothesis is that ML is the most accurate way to predict substrate specificity, someone could ask “How well does this compare to other bioprospecting approaches?” Therefore, we could use a more conventional bioprospecting approach to select an additional 30 thioesterases from the *Cuphea* pool to see if we get the same or a different result. There are two approaches I am familiar with, but I could brush up to see if there are other prominent ones (Enzyme Similarity Tool from Gerlt lab and Greedy algorithm to maximize sequence diversity among homologs).

My judgment on this figure 3B is that we should do this if we do in fact identify the C10-specific thioesterase with the ML algorithm. If not, this figure/experiment is less important

References

[1]

Structure-sequence-function study ended up in Nature Comm.??

Identifies positively charged residues which interact with ACP

Names some C12 thioesterase variants in the introduction

Defines subclasses of FatBs

-Important that these subclasses can have 70% seq ID, but still very different specificities

Used domain swapping and site-directed mutagenesis to ID 11 residues important for specificity

[2], [3]

Make claim that primary sequence info cannot be used to predict substrate specificity

[4]

Changes of M230 in Cuphea thioesterases to Iso, Val, Leu, Phe increases activity

Mention expression in E. coli does not translate well to expression in algae

[5]

Chimeras of CpFatB1 and 2 to explore substrate specificity

[6], [1], [7]

Implicates docking pad interaction between ACP and thioesterase

[8]

27 unique sequences from site saturation mutagenesis. Among the 307 residue positions, 65 positions were chosen for random mutagenesis with 2-8 possible substitutions at each position, and consensus sequences were used for other positions.

[9]

First cloning of UcFatB

[10], [4]

Fruitful examples of *Cuphea* thioesterases used for medium-chain FFA production in *E. coli*

[11]

Lists cuphea species with high C10 oils in seed extracts

[1] F. Jing, L. Zhao, M. D. Yandeau-Nelson, and B. J. Nikolau, “Two distinct domains contribute to the substrate acyl chain length selectivity of plant acyl-ACP thioesterase,” *Nat. Commun.*, vol. 9, no. 1, p. 860, 2018.

[2] F. Jing *et al.*, “Phylogenetic and experimental characterization of an acyl-ACP thioesterase family reveals significant diversity in enzymatic specificity and activity,” pp. 1–16, 2011.

[3] A. Jones, H. M. Davies, and T. A. Voelker, “Palmitoyl-Acyl Carrier Protein ( ACP ) Thioesterase and the Evolutionary Origin of Plant ACyl-ACP Thioesterases,” *Plant Cell*, vol. 7, no. March, pp. 359–371, 1995.

[4] D. Davis, S. Franklin, J. L. Moseley, and R. Bhat, “Variant Thioesterases and Methods of Use,” 10246728, 2019.

[5] M. Ziesack *et al.*, “Chimeric fatty acyl-acyl carrier protein thioesterases provide mechanistic insight into enzyme specificity and expression,” *Appl. Environ. Microbiol.*, vol. 84, no. 10, pp. 1–12, 2018.

[6] M. J. Serrano-vega, R. Garces, and E. Martinez-Force, “Cloning , characterization and structural model of a FatA-type thioesterase from sunflower seeds ( Helianthus annuus L .),” *Planta*, vol. 221, pp. 868–880, 2005.

[7] S. Sarria, T. G. Bartholow, A. Verga, M. D. Burkart, and P. Peralta-Yahya, “Matching Protein Interfaces for Improved Medium-Chain Fatty Acid Production,” *ACS Synth. Biol.*, vol. 7, no. 5, pp. 1179–1187, 2018.

[8] F. Jing, “Characterization of acyl-ACP thioesterases for the purpose of diversifying fatty acid synthesis pathway,” 2013.

[9] T. A. Voelker and H. M. Davies, “Alteration of the Specificity and Regulation of Fatty Acid Synthesis of Escherichia coli by Expression of a Plant Medium- Chain Acyl-Acyl Carrier Protein Thioesterase,” *J. Bacteriol.*, vol. 176, no. 23, pp. 7320–7327, 1994.

[10] N. J. Hernández Lozada *et al.*, “Highly Active C 8 -Acyl-ACP Thioesterase Variant Isolated by a Synthetic Selection Strategy,” *ACS Synth. Biol.*, vol. 7, no. 9, pp. 2205–2215, 2018.

[11] W. B. Phippen, T. A. Isbell, and M. E. Phippen, “Total seed oil and fatty acid methyl ester contents of Cuphea accessions,” *Ind. Crops Prod.*, vol. 24, no. 1, pp. 52–59, 2006.