Design of a Random Sequence Library for Directed Evolution of Cholesterol Oxidase

Jacob Bland and Taylor Shimono

Introduction

Cholesterol oxidase is a protein product that contributes to the breakdown of cholesterol in the steroid biosynthesis pathway (1). Research has shown that cholesterol oxidase from *Streptomyces* is lethal to boll weevils via cell lysis in the midgut. Expression of the cholesterol oxidase gene in cotton plants has the potential to work as a powerful natural pesticide against boll weevils (2). The creation of a random sequence library of cholesterol oxidase genes can be used for directed evolution in cotton plants, conferring resistance to devastation by boll weevils.

To generate this random sequence library, we propose the use of a multiple sequence alignment (MSA) to identify variable and conserved regions in *Streptomyces* protein sequence. These variable regions will provide a basis for the generation of a random sequence library, as shown below:

```
MSQESSAQSRDVTPEPAA[AT]APPT[AD][GE][GD][GA]YDYDVL[IV]VGSGFGGAVSALRLTEKGYRV
GVLEAGRRFTRA[TS]LPR[NS]SWDL[RK]NYLWAPALGLFGIQRIHLLGNVMVLAGAGVGGGSLNYAN
TLYVPP[EK]PFFTDPQWGAITDWQDELKPYYDQA[QS][RK]MLGVRLNPTTTPSDVHLKAAAEKMG[F
V]GDTFHMAPVGVFFGDGKDA[ET]G[AE]SRA[RK]PGEEV[AP]DPYFGGAGPSRRAC[NS]ECGECMT
GCRHGAKNTLNENYLYLAEKAGATIHPMTSVVAVTEDSRGGFAVRTLPT[DN]A[QK]RKGRGRTFTAR
RVVLAAGTYGTQTLLHRMKDTGLLP[HR]LSDKLGTLTRTNSEGLVGAQTTDRRYRKAHGAEK[AV]D
FTKGVAITSSVHPNDSTHIEPVRYGKKSNAMGGLTILQVPYAENSSRVAGFLANCA[RK]HPLLV[LI]RSL
SNRRWSERTIIGLVMQ[TS]LDNSLTTYRKPKGIGKGLLTARQGHGAPNPTQIKEAGEAATAIA[AE]EING
FAGSNIGEL[MI]GTPLTAHFLGGCPIGATADEGV[IV]DPYHRLYGHPGI[TS]VVDGSAVSANLGVNPSLT
ITAQAERAASLWPNKGEADPRP[AE]QGHPYERIAPVAPLRPAVPAEAFAAL
```

This sequence is 600 residues long, with 30 variable sites, denoted by brackets. [DN], for example, means that either aspartic acid or aspargine could be incorporated. It is this random choice that allows the directed evolution to explore the state space for a more optimal sequence.

Design

In the generation of our random sequence library, we began by finding the protein sequence for a *Streptomyces* cholesterol oxidase protein. The original sequence that we chose was from *Streptomyces albidoflavus*, with the sequence information taken from NCBI (3). This was chosen because it is from a strain of *Streptomyces*, which Purcell et al. mentioned in their original article. As with the *Streptomyces* strain described in the paper, this is a type I cholesterol oxidase, which uses FAD as a cofactor to non-covalently bind cholesterol to the protein (4).

(1) UniProt Consortium. (2021, September 29). *Cholesterol oxidase*. UniProt. Retrieved December 8, 2021, from https://www.uniprot.org/uniprot/P9WMV9. (2) Purcell, J. P., Greenplate, J. T., Jennings, M. G., Ryerse, J. S., Pershing, J. C., Sims, S. R., Prinsen, M. J., Corbin, D. R., Tran, M., Sammons, R. D., & Stonard, R. J. (1993). Cholesterol oxidase: A potent insecticidal protein active against boll weevil larvae. *Biochemical and Biophysical Research Communications*, *196*(3), 1406–1413. https://doi.org/10.1006/bbrc.1993.2409. (3) U.S. National Library of Medicine. (2018). *Cholesterol oxidase [streptomyces albidoflavus] - protein - NCBI*. National Center for Biotechnology Information. Retrieved December 8, 2021, from https://www.ncbi.nlm.nih.gov/protein/EFE81671.1. (4) Volontè, F., Pollegioni, L., Molla, G., Frattini, L., Marinelli, F., & Piubelli, L. (2010). Production of recombinant cholesterol oxidase containing covalently bound FAD in Escherichia coli. BMC biotechnology, 10, 33. https://doi.org/10.1186/1472-6750-10-33

After selection of the original protein, we used a multiple sequence alignment to compare this protein to other forms of *Streptomyces* cholesterol oxidase proteins. These were all type I cholesterol oxidases that were very similar to the original protein that was selected. In particular, there was a 97% sequence similarity between the two sequences that were specifically from *Streptomyces*. These sequences then shared an average 86% similarity with the third sequence. We chose to use similar, highly-conserved proteins to ensure that our random sequence library would be relatively similar to native *Streptomyces* proteins. Although this would not necessarily provide a broad range of functions in our novel proteins, it increases the probability that the proteins would fold properly and be viable when encoded by another species, such as the cotton plant. The findings of Purcell et al. suggest that *Streptomyces* cholesterol oxidases are effective against boll weevils, driving us to create a random library that will optimize an already-functional protein.

As expected, our multiple sequence alignments showed a high rate of conservation across the proteins from various *Streptomyces* strains. With this information, we chose the sites with the highest variability to create our sequence library. This procedure allows us to ensure that the regions of the cholesterol oxidase protein that provide crucial function are present in each of our proteins, while making minor modifications to the variable regions that should have small but important effects on protein function. These modifications might change the optimal pH and temperature at which the enzyme functions, but should not remove its ability to lyse midgut cells in boll weevils. This allows for screening of optimal functionality when inserted into other plants, like cotton, via genetic engineering.

Methodology

Please refer to the Jupyter notebook for full visibility into our implementation. While a couple of libraries were considered for other multiple sequence alignment -- biopython, blast -- muscle proved to be easy to use and fast. In fact, with time complexity of O($N^2L + NL^2$), with an additional O(N^3L) for a more refined term, this proved to be not a problem with our N=3 and L=600. With a multiple sequence alignment in hand, we step through the first 600 positions of each aligned sequence. If a given aligned sequence has a legitimate residue, as opposed to a gap, at that position, we add that residue to a set of possible residues for that position. After constructing these sets, we walk through them, constructing a final sequence that has at least one and at most two residues chosen from each positional set.

(1) UniProt Consortium. (2021, September 29). *Cholesterol oxidase*. UniProt. Retrieved December 8, 2021, from https://www.uniprot.org/uniprot/P9WMV9. (2) Purcell, J. P., Greenplate, J. T., Jennings, M. G., Ryerse, J. S., Pershing, J. C., Sims, S. R., Prinsen, M. J., Corbin, D. R., Tran, M., Sammons, R. D., & Stonard, R. J. (1993). Cholesterol oxidase: A potent insecticidal protein active against boll weevil larvae. *Biochemical and Biophysical Research Communications*, *196*(3), 1406–1413. https://doi.org/10.1006/bbrc.1993.2409. (3) U.S. National Library of Medicine. (2018). *Cholesterol oxidase [streptomyces albidoflavus] - protein - NCBI*. National Center for Biotechnology Information. Retrieved December 8, 2021, from https://www.ncbi.nlm.nih.gov/protein/EFE81671.1. (4) Volontè, F., Pollegioni, L., Molla, G., Frattini, L., Marinelli, F., & Piubelli, L. (2010). Production of recombinant cholesterol oxidase containing covalently bound FAD in Escherichia coli. BMC biotechnology, 10, 33. https://doi.org/10.1186/1472-6750-10-33