Design of an Emulsified Protein Binding System using Classification Algorithms

Rylan Marianchuk, Juan Alvarez, Jake Liu, Marija Drikic, Elke Lohmeier-Vogel, Mayi Arcellana-Panlilio

UNIVERSITY OF CALGARY

iGEM 2019 Calgary

Problem Outline

Canada annually exports 9 billion dollars worth of canola seed, oil, and meal to countries such as USA, China, Japan, and Mexico. If canola is exposed to frost before harvest, chlorophyll degredation is stopped, resulting in green seed (figure 1). Oil produced from *green seed* retains chlorophyll and oxidizes quickly, resulting in millions of monetary losses for farmers.

Solution

Emulsify water soluble chlorophyll binding proteins to remove chlorophyll from canola oil at the processing level.

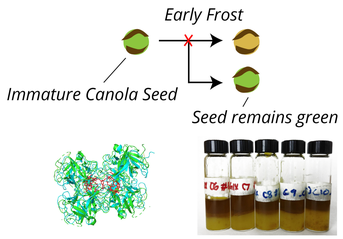
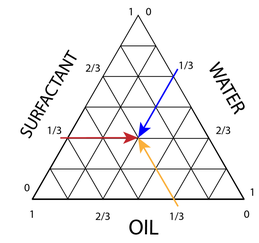


Figure 1: The issue of canola seeds retaining chlorophyll (top), and chlorophyll binding protein, coined 6GIX (left) used in the aqueous phase of various emulsifications (right).

Research Question

How can classification algorithms find the best formulation of an emulsion to optimize chlorophyll extraction from green seed canola oil?

Legend



Background

Canola oil, food grade surfactant, and aqueous buffered saline containing chlorophyll binding proteins were mixed in different ratios and the resulting phase equilibria (figure 2) were observed in order to generate training data for the mathematical model (figure 3).

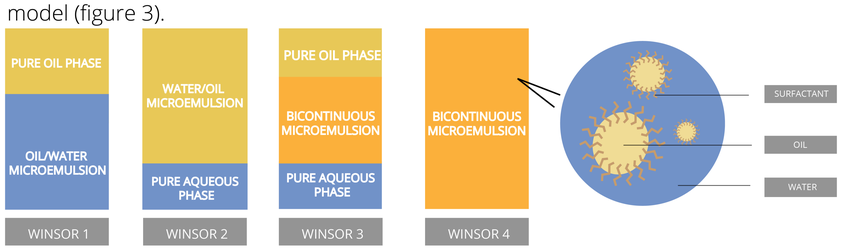


Figure 2: The various Winsor classes of emulsion equilibria containing surfactant, an oil and aqueous protein phase. The classes are dependent upon the ratio of its chemical constituents.

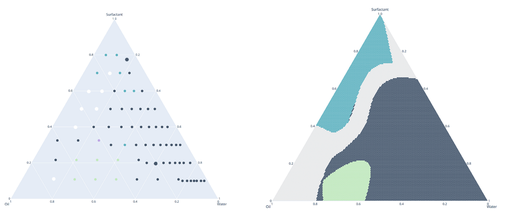


Figure 3: The approach at applying machine learning classification from data observed experimentally to determine the phase class for all possible emulsion compositions. See legend for ternary coordinate structure. Color represents the various Winsor phase classes.

Ultimately use the model to find emulsion compositions that have:

A. Winsor 1 Classification
B. Classifiation confidence equal 1
C. Water to oil ratio less than 1 and greater than 1/3

OPTIMAL

Deem the emulsions satifying A, B, C as optimal, and implement them with chlorophyll binding proteins to extract chlorophyll from canola oil.

Theory and Experimental Results

The two athgorithms we used to classify all possible emulsion compositions are: (1) k-Nearest Neighbours (k-NN) and (2) a Support Vector Machine (SVM), both used the training data collected experimentally.

Both approaches calculate the objective function $\mathcal{F} : \vec{x} \longrightarrow y$ for a classification problem given that:

$$\vec{x} = \begin{bmatrix} r_{oil} \\ r_{water} \\ r_{surfac} \end{bmatrix}, \quad r_{oil} + r_{water} + r_{surfac} = 1, \quad y \in \{Winsor1, Winsor2, Winsor3, Winsor4\}$$

Where $\vec{x}$ represents an emulsion formulation and y is the Winsor class of that formulation.

(1) k-Nearest Neighbours Classification

k- Nearest Neighbours approaches classification by looking at the k nearest features in the training set to categorize the point $\vec{x}$ in question. It can be summarized by computing a probability for each class y:

$$Pr(Y = y \mid \vec{x}) = \frac{1}{k} \sum_{\vec{x}_i \in N_k(x)} I(y_i = y)$$

where $N_k(x)$ is a set containing the k nearest points to $\vec{x}$ and I is the identity function, outputting 1 if true and 0 otherwise. The choice for the label is the y with the maximum probability (Hastie et al. 2017).

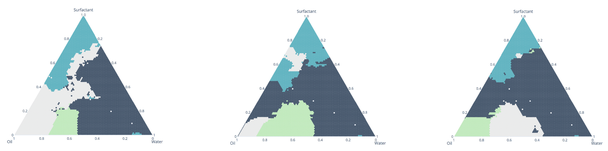
	WINSOR 1	WINSOR 2	WINSOR 3	WINSOR 4
Temperature	300 K		310 K	323 K
k-NN				
k	4		5	5
Error Rate	0.3428		0.2428	0.2142

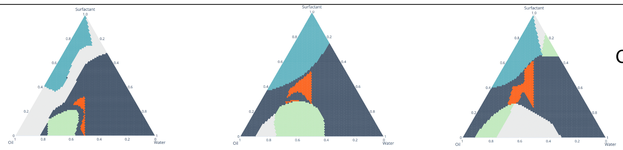
Figure 4: (Above) The best performing models when k-Nearest Neighbours (k-NN) is applied on the training data. The k chosen had the lowest error rate over all other values. The error rate represents the probabilty any given point is classified incorrectly. No optimal emulsions were identified since SVM had a greater model performance. Legend shows the coordinate structure of ternary plots.

(2) Support Vector Machine

Support Vector Machines (SVM) construct a dividing hyperplane for classification defined by the vector $\beta \in \mathbb{R}^3$ and solves the optimization problem:

$$min \ ||\beta|| \quad subject\ to \quad \begin{matrix} y_i(x_i^T\beta + \beta_0) \geq 1 - \epsilon_i \ \forall_i \\ \epsilon_i \geq 0, \quad \sum \epsilon_i \leq \mathcal{C} \end{matrix}$$

where $\sum \epsilon_i$ is the total distance of points on the wrong side of the margin, constrained by $\mathcal{C}$ which is subject to choice (Hastie et al. 2017).

Temperature	300 K	310 K	323 K
SVM			
Gamma	0.72	0.25	0.75
Cost	9400	12200	8800
Error Rate	0.2428	0.1714	0.1571

OPTIMAL

Figure 5: (Above) The best performing models when searching the parameter space on with a Support Vector Machine (SVM) using a non-linear radial basis function kernel with gamma and cost as variables. The error rate represents the probabilty any given point is classified incorrectly. Red points on SVM indicate the emulsion points that satisfy the optimal criteria. Legend shows the coordinate structure of ternary plots (see left for colors).

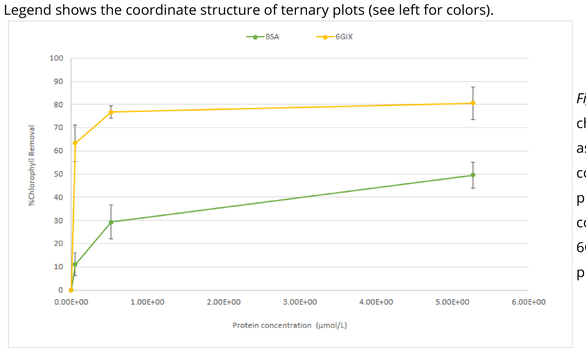


Figure 6: (Left) The change in chlorophyll content in the oil phase as a function of protein concentration within the aqueous phase, using one optimal emulsion composition found by the SVM model. 6GIX is the chlorophyll binding protein. BSA is a control protein.

Future Directions

Figure 6 demonstrates that percent chlorophyll removed from canola oil plateaus with increasing 6GIX protein concentration in the aqueous phase of the emulsion, suggesting that structural instability of the protein is affecting its ligand binding capacity. Future plans hope to engineer a more stable chlorophyll binding protein.

References

A Winsor, B. P., & Hahn, von. (1932). HYDROTROPY, SOLUBILISATION AND RELATED EMULSIFICATION PROCESSES. PART I. Aqueous Solutions of Parajin Chain Salts (Vol. 62). Retrieved from https://pubs-rsc-org.ezproxy.lib.ucalgary.ca/en/content/articlepdf/1948/tf/tf9484400376.

Hastie, T., Friedman, J., & Tisbshirani, R. (2017). The Elements of statistical learning: Data mining, inference, and prediction. Springer.

ATTRIBUTIONS

Thank you to Cassandra Sillner, Sara Far, Nimaya De Silva, Sravya Kakumanu, Christian Emond, Michaella Atienza, Michael Wahba, Jean-David Rousseau, Andrew Symes, Sarah Walker, Prabhkeerat Dhatt, and the iGEM Calgary team for supporting this research. Thank you to Dr. Marcus Samuel and Dr. Nashaat Nassar from the University of Calgary for informing our wet lab project design. Thank you to Dallas Gade and Laurence Parslow from Richardson Oil Seed for being key industry contacts. Thanks to Jacob Grainger, Patrick Wu, Emily Hicks, and Robert Mayall for being mentors to the iGEM Calgary team. Thank you to Dr. Sonny Chan for help attaining funding for this research through the Program for Undergraduate Research Experience.