TOXIN DIAGNOSTICS – DEVELOPMENT OF NOVEL, FIELDABLE TECHNOLOGIES TO DIAGNOSE TOXIN EXPOSURE

Modeling Osmolyte Effects On Protein Toxin Structural Stability

Jaime Rodriguez Pacific Northwest National Laboratory **Nathan Baker** Pacific Northwest National Laboratory

This work aims to leverage machine learning tools to predict the influence of osmolyte species on protein toxin stabilization. Many protein toxins, including but not limited to ricin and abrin, have emerged as possible biological weapons. When biological samples such as these are encountered in the field, preservation is necessary to ensure the sample arrives in an adequate state to allow for rapid characterization and diagnostics by warfighters. Traditional stabilization methods result in modification of the toxin structure itself which is undesirable. Organisms in nature have adapted by using small organic molecules called osmolytes to maintain biomolecular stability. Applying this strategy to protein toxins in liquid matrices has shown to stabilize the native state, and therefore maintain biological activity. Previously, models developed by Auton and Bolen have been capable of predicting the influence of osmolytes on protein stability based on the free cycle of transitions between the protein's native and denatured states. This influence is measured by the m-value, in which a negative m-value denotes a denaturing effect on the protein and a positive m-value denotes a stabilizing effect. However, challenges occur when attempting to extend the model to additional osmolytes, due to the need to obtain experimental data which is costly and time consuming. We propose the use of statistical models and quantitative structure-property relationships (QSPR) as an alternative method to expand the current model. A dataset consisting of experimental m-values for 71 unique osmolyte/protein combinations was procured for initial model development. Molecular descriptors, which are mathematical representations of various molecular properties, were calculated as input features for the 9 osmolytes present in the dataset using the open-source cheminformatics software Mordred. In total, approximately 800 descriptors were calculated for each osmolyte. In addition, solvent-accessible surface areas (SASA) for the backbone and sidechains of each protein residue type were also used as input features. A Lasso regression approach was selected for model development due to the ability to perform feature selection to interpret results. An 80/20 testing and training data split was used with 5-fold cross validation to tune parameters. Preliminary results showed an r squared value of 0.828 along with root mean squared and mean absolute errors of 948.4 and 653.9 cal/mol/M, respectively. Qualitatively, the model correctly predicted the stabilizing and denaturing effects of each osmolyte. The robustness of the model was evaluated by leaving each osmolyte out of training and testing and then introducing afterwards in a validation step. The model was found to be significantly influenced by the exclusion of certain osmolytes from the training and testing portions. Work in progress is focused on improving model robustness by utilizing an expanded dataset of 380,000 osmolyte/protein combinations with m-values calculated from the Auton and Bolen model along with the inclusion of additional input features obtained by deep learning models.