PREDICCIÓN DE ESTABILIDAD DE PROTEÍNAS MUTANTES USANDO VECTORES DE AUTOCORRELACIÓN DE SECUENCIA DE AMINOÁCIDOS (AASA) Y MACHINE LEARNING

BENJAMÍN GUSTAVO PINO RAMÍREZ
INGENIERO CIVIL EN COMPUTACIÓN

RESUMEN

Contexto: Las proteínas son moléculas con una gran diversidad de funciones en la naturaleza. Estas poseen una propiedad medible, que es la estabilidad conformacional, la cual se relaciona con su resistencia a altas temperaturas. Proteínas con alta estabilidad tienen muchas aplicaciones, pero desarrollarlas en un proceso costoso. Para asistir al experto en este proceso, se ha estudiado el uso de técnicas de machine learning para la predicción de estabilidad de proteínas mutantes. Los vectores AASA son una representación cuantitativa de las proteínas que ha demostrado ser útil para modelar la estabilidad en investigaciones anteriores. Sin embargo, solo se ha aplicado en conjunto con técnicas de modelado no lineal. Si modelos más simples, como lo son los lineales, tienen un desempeño de predicción bueno o, al menos, similar al de modelos más complejos, los primeros son más deseables pues son más interpretables y útiles para análisis posteriores. Problema: Desarrollar un método para modelar la estabilidad de las mutantes y que permita comparar modelos lineales y no lineales. Solución propuesta: Se propone una metodología para determinar, empíricamente, si es que las técnicas de modelado lineal tienen un buen desempeño para la predicción de la estabilidad a partir de vectores AASA, y cómo su desempeño se compara con el de técnicas de modelado no lineal. Esta metodología es aplicada a cuatro conjuntos de datos distintos, evaluando y comparando el desempeño de cuatro técnicas de modelado lineal y una de modelado no lineal, usando tres variantes de Cross-validation (CV), Nested CV, 5-Fold CV y 5x2 CV. Resultados: Se observa que 5-Fold CV y 5x2 CV producen estimaciones de desempeño con menos variabilidad que Nested CV, por lo que son más fiables. Se observa, además, que el desempeño de las técnicas de modelado lineal es consistentemente bajo a través de los distintos conjuntos de datos abordados y,

en la mayoría de los casos, inferior al desempeño de la técnica de modelado no lineal. Conclusiones: Se observa una superioridad de los métodos no lineales, en particular de SVR con kernel RBF, en el reconocimiento de patrones en los datos. Adicionalmente, se obtiene una metodología reproducible, útil para el análisis de conjuntos de datos disponibles a futuro, ya que es independiente de las técnicas de modelado y la implementación realizada en este trabajo es de libre acceso.