A Machine Learning Approach to Identifying Ordered Binding Regions on Order-Disorder Protein Interfaces

EXECUTIVE SUMMARY

If a cell is thought of as an intricate machine, proteins are the gears that connect individual components and allow for overall cellular function. All proteins are made of linked sequences of smaller compounds known as amino acids. Scientists once believed that all proteins had only one rigid structure based off their amino acid sequence under the "lock and key" model of protein interaction. Under this model, only one specific protein can act as a "key" to interact with another protein, the "lock." In the past few decades, scientists have found this model to be incomplete. There are flexible proteins that can change their 3D structures to bind to various partners. These regions are called intrinsically disordered regions (IDRs) and have been found in various disease causing proteins. IDRs are hard to target with traditional methods due to their dynamic properties. The goal of this project was to create a program that can identify specific binding sites on proteins that are known to interact with IDRs. A computational technique called *machine learning* was implemented where the program is able to train itself and learn. The generated algorithm was able to predict disordered binding sites with approximately 76% accuracy. With further improvements to a program of this nature, drugs that bind to disordered protein sites can be developed. Additional drugs that affect protein interactions can be easily developed to treat many life threatening diseases such as cancer, Ebola, and cardiovascular disease.

A Machine Learning Approach to Identifying Ordered Binding Regions on Order-Disorder Protein Interfaces

ABSTRACT

Intrinsically disordered proteins (IDPs) and intrinsically disordered regions (IDRs) are areas on proteins that lack rigidly defined three-dimensional structures. Due to their conformational flexibility, IDRs are commonly involved in cellular signaling and regulation than ordered proteins. Unsurprisingly, many proteins important in disease, including BRCA1 for breast cancer or tau for Alzheimer's disease, possess significant regions of protein disorder. Although there are existing algorithms that are capable of predicting IDR protein binding partners, there is still no way to predict specific binding regions on such proteins. Here, a machine learning approach was implemented to generate a Bayesian network that models relationships between protein characteristics. A sliding window algorithm was created with a binary classifier that can identify the presence of binding sites on 10-residue ordered protein segments at approximately 76% accuracy. These results provide the groundwork for a comprehensive IDP binding-site identifier that will allow for the creation of modern drug treatments targeting regions of intrinsic disorder.