Statistical Learning of Binding Activity Using Structural Characteristics of Protein Binding Sites

In order to improve our understanding of functions of cells or organisms, it is critical to learn about the functions of individual proteins, which are determined by their structures and are tied to the molecules the proteins bind to. Recently, many projects have aimed to find representative proteins for each protein family ([1], [2], [3]) to deduce the functions of structurally similar proteins via mathematical and statistical methods. Many researchers have focused specifically on studying binding sites, which are comprised of atoms near the surfaces of protein molecules where binding activity is known to occur.

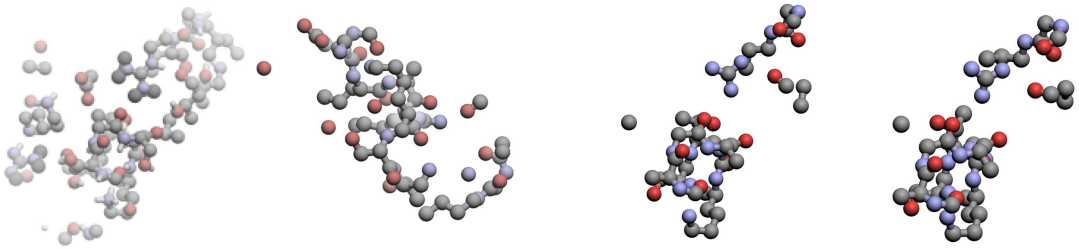

Figure 1: Left: two unaligned binding sites. Right: aligned common atoms

In the left side of Figure 1, two binding sites that bind to the same molecule are shown. It is difficult to compare these sites directly because the atoms are shown in different coordinate systems. A common approach for solving this problem is to align the binding sites in a common coordinate system and then compare characteristics of the sites to each other [4]. On the right side of Figure 1, the atoms common to both of the binding sites are shown with respect to the same coordinate system. Unfortunately, this approach is computationally challenging and produces similarity scores that are difficult to use to develop statistical tools to both model groups of similar binding sites and predict binding activity of new sites. However, such models and classification studies are key to help researchers develop improved understanding of the relationship between protein structures and their binding activity.

In this project, we will instead build upon a recent paper [5] that introduced a novel representation of a binding site as a covariance matrix summarizing the structural information in a way that bypasses the coordinate system problem. This will allow us to utilize tools from multivariate analysis to develop statistical models for groups of binding sites. We will utilize these models to better understand both the notion of binding sites from a mathematical perspective and the relationship between the structure and binding activity of binding sites. As part of this, we will explore a variety of statistical machine learning methods to perform classification studies.

Timeline: Week 1: Students will learn about background information on proteins, multivariate statistical analysis, and many commonly used classification algorithms, such as nearest neighbor methods, logistic regression, classification trees, and random forests. Weeks 2-3: Students will work with binding sites from benchmark data sets to get experience working with the data and also implement a variety of classification methods for a number of other types of data to familiarize themselves with them and learn their various strengths and weaknesses. Weeks 4-6: Participants will implement the classification procedures for predicting protein binding activity and aggregate the results. Weeks 7-8: Participants will draw conclusions about which categories of binding sites the classification procedures work well for and which ones they do not to identify areas for further research and write up the paper.

References

[1] Burley SK (2000) An overview of structural genomics. Nat Struct Biol 7:

[2] Stevens RC, Yokoyama S, Wilson IA (2001) Global efforts in structural genomics. Science 294: 8992.

[3] Montelione GT (2001) Structural genomics: an approach to the protein folding problem. Proc Natl Acad Sci U S A 98: 1348813489.

[4] Ellingson, L. and Zhang, J. (2012). Protein Surface Matching by Incorporating Local and Global Geometric Information. PLoS ONE 7(7): e40540. doi:10.1371/journal.pone.0040540

[5] Premarathna, G.I. and Ellingson, L. (2021) A mathematical representation of protein binding sites using structural dispersion of atoms from principal axes for classification of binding ligands. PLOS ONE 16(4): e0244905. https://doi.org/10.1371/journal.pone.0244905