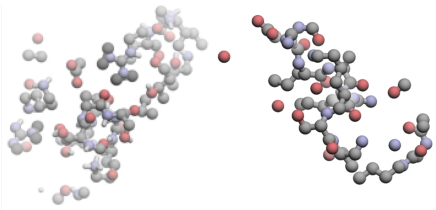
Statistical Modeling of Variation in Structural and Chemical Characteristics of Protein Binding Sites

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.

One approach established by a recent paper [4] and further studied in last year's REU [5] utilized the inherent multivariate nature of the data to study the data. [4] developed a novel encoding of a binding site's structural information as a covariance matrix and [5] further incorporated the chemical composition of the binding sites. Both studies illustrated that this representation of the data is is highly effective for classifying binding sites according to the ligand they bind to.

In this project, we will build off of [4] and [5] to gain new insights into how this representation can help us understand binding sites, including the very definition of them. Binding sites are typically defined to consist of all atoms in a protein within a specified distance to the ligand it binds to, but there is not a consensus on exactly what that distance should be. We will investigate how the choice of this distance affects both the distributions, including the means and covariances, of binding sites within each group and the performance of the classification techniques utilized in [4] and [5].



$$\begin{pmatrix} 2.75 & -1.16 & 0.05 \\ -1.16 & 8.17 & 6.82 \\ 0.05 & 6.82 & 9.46 \end{pmatrix}, \begin{pmatrix} 3.34 & -1.12 & 0.21 \\ -1.12 & 8.61 & 6.18 \\ 0.21 & 6.18 & 5.84 \end{pmatrix}$$

Figure 1: Left: two binding sites. Right: their covariance matrix representations

Timeline: Week 1: Students will learn about background information on proteins, multivariate statistical analysis, and many commonly used classification algorithms. Weeks 2-3: Students will work with binding sites from benchmark data sets to explore the data and perform a literature review to understand how binding sites are commonly chosen. Weeks 4-6: Participants will investigate the impact size of binding sites on distributions of the data and how these affect classification studies. Weeks 7-8: Students will draw conclusions about what they have found through their studies, write up a written report, and prepare a final oral presentation and poster presentation.

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] 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

[5] Martirosyan, V., Smith, E., Hill, K., and Ellingson, L. (2022) Classification of protein binding sites using structural and chemical information. Poster presented on July 29, 2022.