5. 考察

5.1 クリプトサイトになり得る凹みとその他の凹みのデータ分布比較についての

考察

　図　と表　より、クリプトサイトになり得る凹みとその他の凹みについて、Fpocket[4]より得られた各特徴量の内、Hydrophobicity score、Mean alp. sph. solvent accessがP値 < 0.05であり、データ分布に有意に差があることがわかる。これから、疎水性や溶媒接触面積がクリプトサイトになり得る凹みとその他の凹みの違いを生み出す因子として考えられる。なお、Hydrophobicityの結果については先行研究[1]の考察と一致する。

5.2 クリプトサイトになり得る凹みとその他の凹みの主成分分析(PCA) についての

考察

図　右上より、学習データの分布内にテストデータの分布が存在し、テストデータも点在していることから、データセットの質に特に問題がないと考えられる。また図　左上より、クリプトサイトになり得る凹みとその他の凹みについて、PC1とPC2のみでは分布に違いが特にない。さらに、図　左下からPC1とPC2のみの累積寄与率は40%未満であり、PC1とPC2のみでは、識別が困難であると考えられる。さらに図　右下より、PC1とPC2が特徴量をどういう観点で評価しているかは特にわからなかった。

5.3 機械学習モデルを用いたクリプトサイト有無の学習と推論に対する性能比較についての考察

表　からSVMが一番性能がよく、テストデータについてF1値で約71.0%の性能を達成した。

5.4 機械学習モデルの予測が正答、誤答した場合のタンパク質ポケット周辺の表面構造の観察についての考察

図 〜から、現状としてモデルが予測を誤答する場合はアポ構造においてクリプトサイトになりうる凹みが浅く、その他の凹みと判定を誤ったと考えられる。

5.5 機械学習モデルの因子分析についての考察

　図　左はデータ全体についての因子分析である。特徴量の内、Hydrophobicity score、Alpha sphere density、Polarity scoreが重要特徴量の上位であることがわかる。また図　右より、モデルの出力に対して、Hydrophobicity scoreとPolarity scoreは正の相関があり、Alpha sphere densityは負の相関があることがわかる。つまりは、Hydrophobicity scoreとPolarity scoreの値が正の値であればあるほど、Alpha sphere densityの値が負であればあるほど、モデルは１(クリプトサイトになり得る凹み)を出力しやすくなるということが考えられる。実際にこの考察は図　の統計分析の結果とよく合致することが確認できる。

　さらに、図　は個々のデータについての因子分析である。予測を高確率(0.8以上)で正答した場合(図　左下)は、Alpha sphere density、Hydrophobicity score、Apolar SASAがモデルの出力に相対的に大きな影響を与えていることがわかる。一方で、予測を誤答した場合(図　右下)は、クリプトサイトになり得るポケットと予測した場合には、同様にAlpha sphere density、Hydrophobicity score、Apolar SASAがモデルの出力に相対的に大きな影響を与えているが、その他のポケットと予測した場合には、特定の特徴が相対的に大きな影響を及ぼしたわけではないことがわかる。それぞれのケースの実際のタンパク質の考察は5.4の通りである。