Abstract

Proteins, a diverse set of macromolecules, are responsible for almost all activities in the simplest of viruses to cells in evolved multicellular organisms. Alzheimer's disease, Parkinson's disease, Type 2 diabetes, Huntington's disease, Cystic Fibrosis, Sickle cell disease, and a number of different diseases occur either due to the toxicity or functional aberration arising out of proteins that fail to fold in their native configurations. Computational methods, by engineering novel proteins, have assisted in the better understanding and treatment of disease conditions.

Zaire ebolavirus (EBOV) viral proteins (VPs) interact with numerous human proteins to completely disrupt both the innate and adaptive immunity in the host system. Among the VPs, the EBOV VP35 protein interacts with the maximum number of human proteins. We consult available structural information to model a tetrameric assembly of the EBOV-VP35 protein and identify inter-chain bonding networks. We use molecular dynamics (MD) simulations along with normal mode analysis to determine the flexibility and deformity in various regions of the protein. We further propose a computational docking- design based protocol to identify critical interaction residues in host-pathogen protein-protein interactions and in process establish that the VP35 protein directly interacts with human PKR to prevent it from autophosphorylation. Our findings can be considered for the shortlisting of potential small molecule inhibitors.

Realizing the importance of protein design in computational protein engineering, we adopt a stochastic approach to *de novo* protein design and primarily focus on the development of an efficient search strategy for optimal amino acid sequences. We present a greedy replica-exchange Monte Carlo search algorithm to ensure faster convergence (6.16× overall speedup) in protein design. We prepare an evolutionary mutation profile of the structural homologs to allow amino acid variability in each residue position of the target protein and ensure the dynamic stopping of simulation trajectories in a stagnant condition. On a benchmark dataset of 76 proteins, our algorithm reports an average root-mean-square deviation of 1.21Å between the target and the design sequences when modeled with protein folding software.

The encoding of nonsynonymous single nucleotide polymorphisms (nsSNP) cannot be addressed by *de novo* protein design. Therefore, we develop ProTSPoM that uses random forest regressors and gradient boosting regressors to encode the change in Gibbs free energy ($\Delta\Delta G$) resulting from single point mutations. ProTSPoM outperforms all existing methods on the S2648 and S1925

databases and reports a Pearson correlation coefficient of 0.82 (0.88) and a root-mean-squared-error of 0.92 (1.06) kcal/mol between the predicted and experimental $\Delta\Delta G$ values on the long-established S350 (tumor suppressor p53 protein) dataset. ProTSPoM identifies SNPs in the DNA binding domain of the p53 protein which are plausibly detrimental to its structural integrity and interaction affinity with the DNA molecule. ProTSPoM, with its reliability and efficiency, can be integrated with existing protein engineering methods.

After point mutations, InDels are the second most frequent form of protein modifications that can remarkably alter the structural and functional specification. We introduce a dataset listing 162 single-point deletions (SPDs) instances (132 lead to folded conformations, rest to unfolded states). Using this dataset we construct a random forest classifier and an elliptic envelope based outlier detector encoding simple structural and physicochemical features to predict the change in foldability arising from SPD instances. Adhering to leave one out cross-validation, the random forest classifier and the elliptic envelope report an accuracy of 99.4% and 98.1% respectively. Further, we present a positive unlabeled learning-based prediction system (PROFOUND) to predict the change in foldability arising from multi-point deletions (MPDs) spanning multiple secondary structures and multiple solvent accessibility states in protein structures. In the absence of an MPDs dataset, we curate 153 MPD instances that lead to native-like folded structures and 7650 unlabeled MPD instances whose effect on the foldability of the corresponding proteins is unknown. Considering our newly introduced dataset, PROFOUND on 10-fold cross-validation reports a recall of 82.2% (85.2%) and a fallout rate of 14.8% (21.5%) in the protein loop (non-loop) region. The low fallout rate indicates that the foldability in proteins subject to MPDs is not random and requires unique specifications. A first of a kind foldability prediction system owing to MPD instances and the new MPD dataset will support protein engineering endeavors.

Keyword: Protein engineering; protein thermodynamic stability; InDels in proteins; protein stability prediction technique; *Zaire ebolavirus*; Web services.

Name- Anupam Banerjee

Roll Number- 14AT91R12