Prediction of protein stability changes upon single point mutation using Ensemble Learning

Alex D. Camargo, Adriano V. Werhli, Karina S. Machado

Centro de Ciências Computacionais, Universidade Federal do Rio Grande - FURG

The analysis of the destabilizing, neutral or stabilizing impact of single point mutations proteins may be extremely valuable to further refine the relationship between sequence, structure and function of proteins. For this reason, computational tools were developed to predict the impact of point mutations. These models imply in the use of different assumptions about the probabilities of amino acid substitutions. Moreover, these assumptions seek for an approximation of reality supported by better accuracy. The combination of computational and statistical methods with experimental techniques, e.g. deep sequencing and high precision stability measurements, provides many supporting approaches for the protein stability engineering albeit the inherent computational problems. Most computational methods calculate the ΔΔG (free energy difference between a wild type protein and its mutant) which is considered an indicator of the mutation effects. Therefore, when there are different competing approaches to this problem, an effort to determine the most accurate is inevitable. The best approach depends on the available data and prior knowledge of the expert. Thus, the adoption of approaches to produce a final result better than individual results was sought using Ensemble Learning, taking into consideration that the values resulting from its classification can add greater generality by consensus. In doing so this work aims at developing an ensemble method to combine the results of different methods for predicting the impact of point mutation in proteins. At this moment we are considering the tools: I-Mutant, CUPSAT, SDM, mCSM, DUET, iRDP and MAESTRO. The initial proposal uses the plurality vote, popular in such learning as the key factor in the set. The dataset used in the case study came from the selection of experimental data from the biological databases Protherm and Protein Data Bank (PDB) totaling 1775 mutations. In general, the predictions had a good accuracy compared to the experimental values. For example, the tools I-Mutant and CUPSAT obtained accuracy of 75.21% and 65.57% of the predictions, respectively. The result the proposed ensemble (plurality vote) was similar to the best individual method, reaching 73.07% accuracy. As future work we are going to apply different mechanisms of ensemble and we are going to develop a tool based on the proposed method.