Prediction of protein structural features by use of artificial neural networks

In the past decades we have seen an exponential growth of biological sequence data. The cost for DNA sequencing has dropped significantly since the announcement of the first sequenced genome and newly sequenced genomes are published almost every week. Publicly available genetic sequence databases like for example GenBank are increasing considerably in size and GenBank currently contains more than 132 million sequences. Similar the Protein Data Bank currently contains more than 71,000 experimentally determined structures of nucleic acids, proteins and nucleic acid/protein complexes. There is a huge over-representation of DNA sequences when comparing the amount of experimentally verified proteins with the amount of DNA sequences. The academic and industrial research community therefore has to rely on structure predictions instead of waiting for the time consuming experimentally determined structure data. This thesis describes the development of two new tools to study such genetic sequence data. NetSurfP was developed to predict the surface accessibility of amino acids in amino acid sequences. Knowledge of the degree of surface exposure of an amino acid is valuable and has been used to enhance the understanding of a variety of biological problems, including protein-protein interaction, prediction of epitopes and active sites. Following NetSurfP, NetTurnp was developed for the prediction of -turn occurrence. Using secondary structure and surface accessibility predictions from NetSurfP, a better understanding and improvement of the performance for the prediction of -turns was obtained. -turns are very interesting in the way that they are the most abundant type of turn structures, and approximately 25% of all amino acids in protein structures are located in a -turn. In bioinformatics speed and accuracy is an important factor, hence the developed tools are expected to return a result in a rapid and efficient manner. Our way of solving that problem was to pre calculate protein sequence data. Currently, more than 500,000 protein sequences are in the local cache. In relation to surface exposure, a third project dealt with the prediction of discontinuous B-cell epitopes. Here Half Sphere Exposure (HSE) was integrated in an existing prediction method. HSE is a measure of solvent exposure where the upper and lower epitope contacts to a given residue can be weighted differently. The integration of HSE showed to improve previously obtained results. Lastly, I present an attempt to predict the HIV-1 Protease specificity. As the protease is essential for the life cycle of the HIV virus, the protease is of great interest as a target for the rational design of drugs against HIV. We show that it is possible to predict the specificity of the HIV protease with a high performance. In the process we also identified new possible cleavage sites which will further be verified experimentally in the lab. In summary, the thesis presented in this work has greatly contributed to the development of new tools in bioinformatics that will hopefully aid in future scientific discoveries.