ABSTRACT

# Computational Prediction Models for Proteolytic Cleavage and Epitope Identification

Liwen You, Department for Intelligent systems, IDE, Halmstad University, SE-301 18 Halmstad, Sweden and Department of Theoretical Physics, Lund University, Sweden

**Abstract**

The biological functions of proteins depend on their physical interactions with other molecules, such as proteins and peptides. Therefore, modeling the protein-ligand interactions is important for understanding protein functions in different biological processes. We have focused on the cleavage specificities of HIV-1 protease, HCV NS3 protease and caspases on short oligopeptides or in native proteins; the binding affinity of MHC molecules with short oligopeptides and identification of T cell epitopes. We expect that our findings on HIV-1 protease, HCV NS3 protease and caspases generalize to other proteases.

In this thesis, we have performed analysis on these interactions from different perspectives --- we have extended and collected new substrate data sets; used and compared different prediction methods (e.g. linear support vector machines, neural networks, OSRE method, rough set theory and Gaussian processes) to understand the underlying interaction problems; suggested new methods (i.e. a hierarchical method and Gaussian processes with test reject method) to improve predictions; and extracted cleavage rules for protease cleavage specificities.

From our studies, we have extended oligopeptide substrate data sets and collected native protein substrates for HIV-1 protease, and a new oligopeptide substrate data set for HCV protease. We have shown that all current HIV-1 protease oligopeptide substrate data sets and our HCV data set are linearly separable; for HIV-1 protease, size and hydrophobicity are two important physicochemical properties in the recognition of short oligopeptide substrates to the protease; and linear support vector machine is the state-of-the-art for this protease cleavage prediction problem. Our hierarchical method combining protein secondary structure information and experimental short oligopeptide cleavage information can improve the prediction of HIV-1 protease cleavage sites in native proteins. Our rule extraction method provides simple and accurate cleavage rules with high fidelity for HIV-1 and HCV proteases. For MHC molecules, we showed that high binding affinities are not necessarily correlated to immunogenicity on HLA-restricted peptides. Our test reject method combined with Gaussian processes can simplify experimental design by reducing false positives for detecting potential epitopes in large pathogen genomes.

**Keywords**: binding affinity; caspase; cleavage prediction; cleavage specificity; epitope; false positive; Gaussian process; HCV; hierarchical method; HIV; immunology; MHC; OSRE; protease-peptide interaction; rule extraction; sequence analysis; SVM