Predicting trypsin cleavage sites based on sequence information

Thomas Fannes[1], Elien Vandermarliere[2,3], Leander Schietgat[1], Lennart Martens[2,3], and Jan Ramon[1]

[1]Department of Computer Science, Katholieke Universiteit Leuven, Belgium
[2]Department of Medical Protein Research, VIB, Ghent, Belgium
[3]Department of Biochemistry, Ghent University, Ghent, Belgium

November 29, 2012

Abstract

Trypsin is the most used enzyme in proteomics experiments to convert proteins into peptides as it has a high substrate specificity, it cuts exclusively after arginine and lysine residues. A typical problem is to identify an unknown protein: The protein is cleaved with trypsin and after mass spectrometry, the resulting spectra are compared to theoretical spectra to allow for an identification of the unknown peptides and thus of the unknown protein. The size of the search space is dependent on the number of possible peptides which is quadratic in the number of possible cleavage positions. Accurately predicting cleavage or miscleavage thus reduces the search space.

In our work we use machine learning techniques to learn a model capable of predicting trypsin cleavage based on the primary structure of a protein and a possible cut position in the sequence. We allow a number of tests on the amino acids type and/or their properties within a window around the possible cut position, e.g. "Is there an amino acid with neutral charge two positions after the cut position?" or "Is there a proline within distance one of the cut position?" We learn a random forest, a set of decision trees where each tree is generated by using a random subset of the test set, and the actual prediction is generated by aggregating the predicted values of the trees in the forest.

We compare the random forest with respect to an existing rules set, the so called "Keil rules". The forest was learned on a homogeneous dataset retrieved from PRIDE[1] by selecting all 681 193 examples containing trypsin cleavage information. Evaluated on the PRIDE dataset, it attains an AUROC of 96%, an improvement of 28% with respect to the Keil rules. The two models are also evaluated on three independent datasets: the iPRG-dataset (9694 examples), the CPTAC-dataset (23842 examples) and the MS_LIMS-dataset (26079 examples). Our method achieves AUROC scores of 84% to 90%, significantly outperforming the Keil rules set

[1]`http://www.ebi.ac.uk/pride/`

with an average improvement in AUROC of 17.9%. We therefore conclude that our trypsin cleavage predictor favorably compares with respect to state-of-the-art models.