Predicting trypsin cleavage sites based on sequence information using decision tree ensembles

Thomas Fannes THOMAS.FANNES@CS.KULEUVEN.BE
Department of Computer Science, Katholieke Universiteit Leuven, Belgium

Elien Vandermarliere ELIEN.VANDERMARLIERE@UGENT.BE
Department of Medical Protein Research, VIB, Ghent, Belgium
Department of Biochemistry, Ghent University, Ghent, Belgium

Leander Schietgat LEANDER.SCHIETGAT@CS.KULEUVEN.BE
Department of Computer Science, Katholieke Universiteit Leuven, Belgium

Lennart Martens LENNART.MARTENS@UGENT.BE
Department of Medical Protein Research, VIB, Ghent, Belgium
Department of Biochemistry, Ghent University, Ghent, Belgium

Jan Ramon JAN.RAMON@CS.KULEUVEN.BE
Department of Computer Science, Katholieke Universiteit Leuven, Belgium

Keywords: decision tree ensemble, proteomics, tryptic cleavage

Proteomics is the large-scale study of proteins, and a typical problem is to identify an unknown protein through mass spectrometry. The protein is cleaved by an enzyme and these peptides are then fed to a mass spectrometer. Afterwards the resulting spectra are compared to in-silico spectra to allow for an identification of the unknown peptides and thus of the unknown protein. Trypsin is the most used enzyme to convert proteins into peptides as it has a high substrate specificity: it cuts exclusively after an arginine and a lysine residue in the protein's sequence.

In our algorithm we propose CP-DT, which is based on a decision tree ensemble and is 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 decision tree ensemble where each tree is generated by using a random selection of tests, and the actual prediction is generated by averaging the predicted values of the trees in the forest. The decision tree ensemble is learned by our in-house MIPS framework, a highly-generic, template-based `C++` data mining tool, capable of handling large data streams.

We compare our model with respect to the state-of-the-art "Keil" rules set. CP-DT was learned on a homogeneous dataset retrieved from all 681 193 examples in PRIDE[1]. The model is evaluated on three independent datasets: iPRG (9694 examples), CPTAC (23 842 examples) and MS-Lims (26 079 examples). CP-DT achieves AUROC scores of 84% to 90%, significantly outperforming the Keil rules set with an average improvement in AUROC of 17.9%. In a final step, we use our model to create a database of peptides by applying Naive Bayes, i.e., the probability of cleavage is the product of the start position's cleavage prediction, the end position's cleavage prediction and no cleavage in the middle. This database is compared to typically used databases where each peptide has at most one miscleavage. Here we achieve an AUROC of 93%, outperforming existing techniques with an improvement of 10.0%, which shows a compression of the tryptic search space with respect to traditional databases. We therefore conclude that our trypsin cleavage predictor outperforms the state-of-the-art model.

Acknowledgments

This research has been supported by ERC Starting Grant 240186 MiGraNT: Mining Graphs and Networks: a Theory-based approach.

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