Predicting the C-Terminal Amino Acid of a Peptide from MS/MS Data Supplementary File 2

Jens Allmer

# Abstract

Proteomics investigates the complement of a genome. It is currently mainly based on mass spectrometry (MS) which is the tool of choice to investigate proteins. Two computational approaches to derive the tandem mass spectrum precursor’s sequence are widely employed. Database search essentially retrieves the sequence by matching the spectrum to all entries in a database whereas *de novo* sequencing does not depend on a sequence database. Both approaches generally benefit from knowledge about the enzyme that was used to generate the measured peptides. Most algorithms default to trypsin since it is abundantly used. Trypsin cuts after arginine and lysine and thus the c-terminal amino acid is not known precisely. Furthermore, 90% of protein terminal peptides may not end with either arginine or lysine and are thus not conformant with the algorithm’s assumptions. Here an algorithm, named RKDecider, to sort the c-terminal amino acid into one of three groups (arginine, lysine, and other) is presented. Although around 90% accuracy was achieved during data mining spectra for rules that determine the c-terminal amino acid, RKDecider’s accuracy is a little less and achieves about 80%.

# Software Implementation of the RKDecider

The software RKDecider is a rule based decision system which based on diagnostic losses of a precursor that can be measured in an MS/MS spectrum. The rules were learned using Orange Canvas [1]. Since accuracy depends on the terminal amino acid and on the charge of the precursor, 6 decision trees were build and then implemented into software. First the three decision trees for deciding whether the c-terminal amino acid is arginine or any other are shown (Figures 1-3) and then the decision trees for lysine are presented (Figures 4-6).

## Predicting C-Terminal Arginine

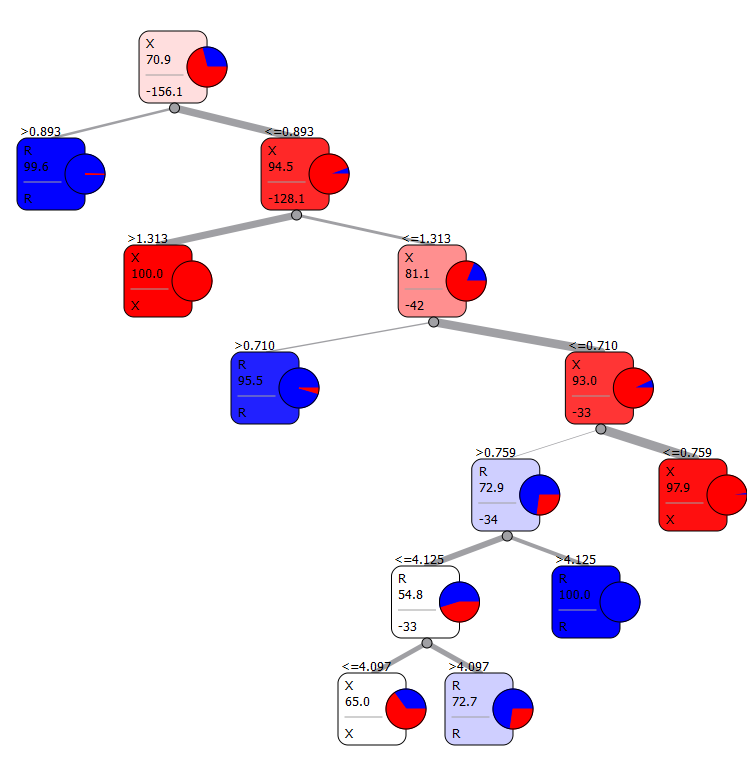


Figure 1: Decision tree to decide whether the terminal amino acid is argining, given a charge of 1 (z=1).

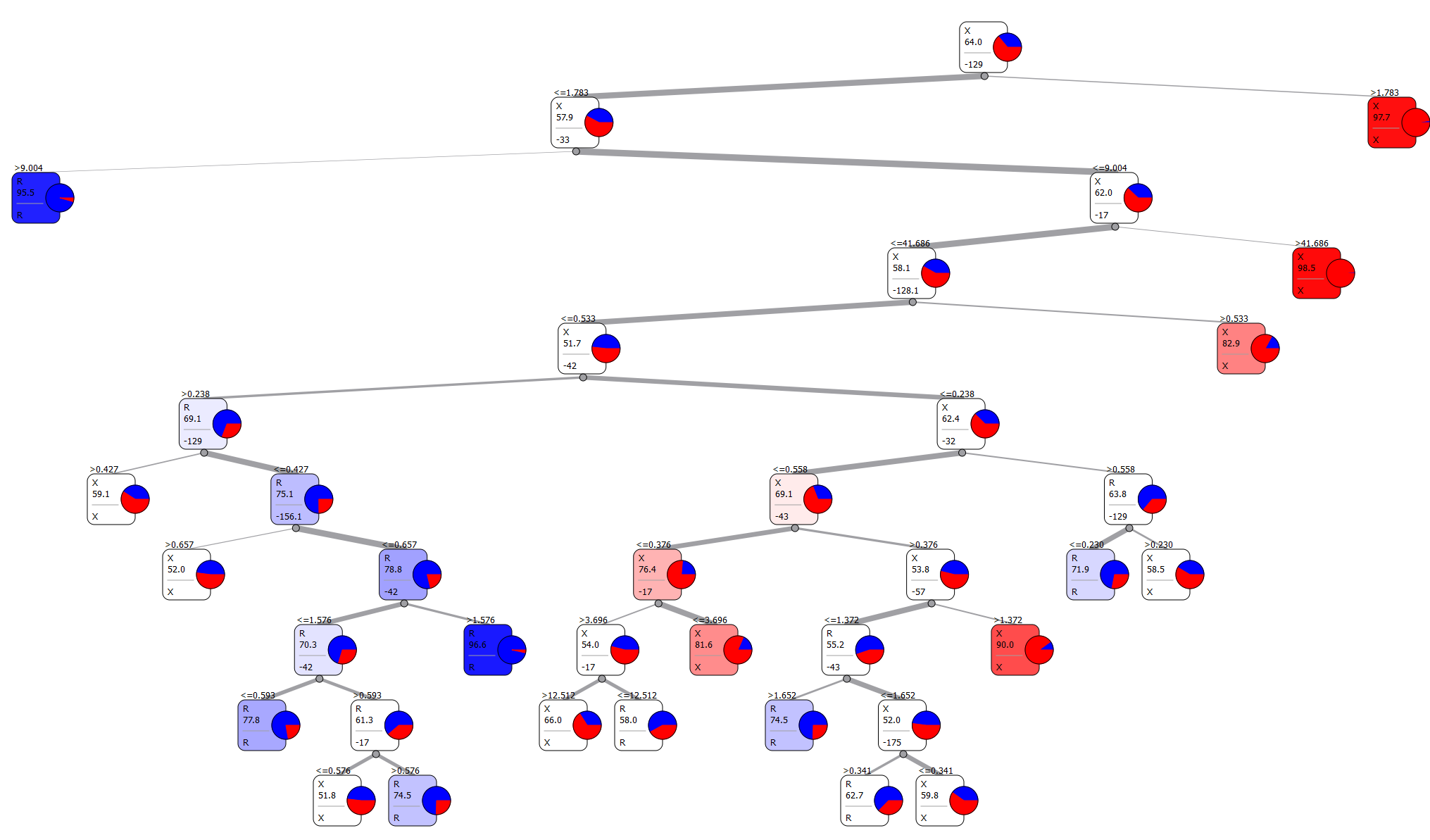


Figure 2: Decision tree to decide whether the terminal amino acid is argining, given a charge of 2 (z=2).

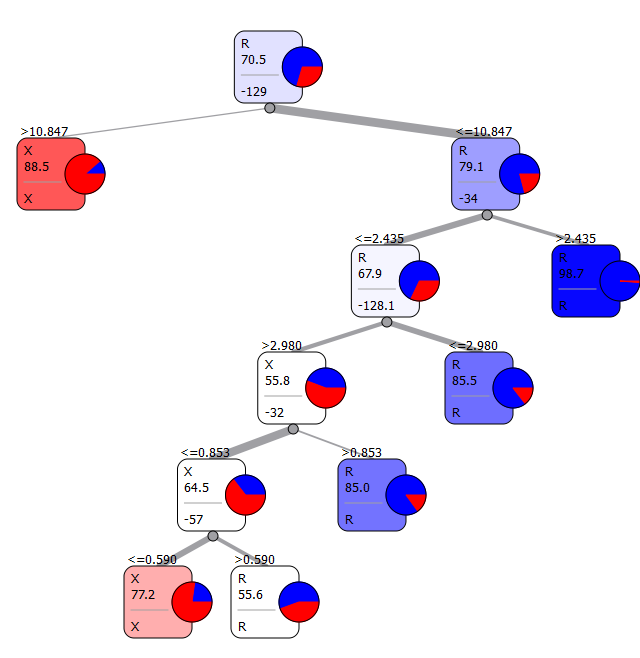


Figure 3: Decision tree to decide whether the terminal amino acid is argining, given a charge of 3 (z=3).

## Predicting C-Terminal Lysine

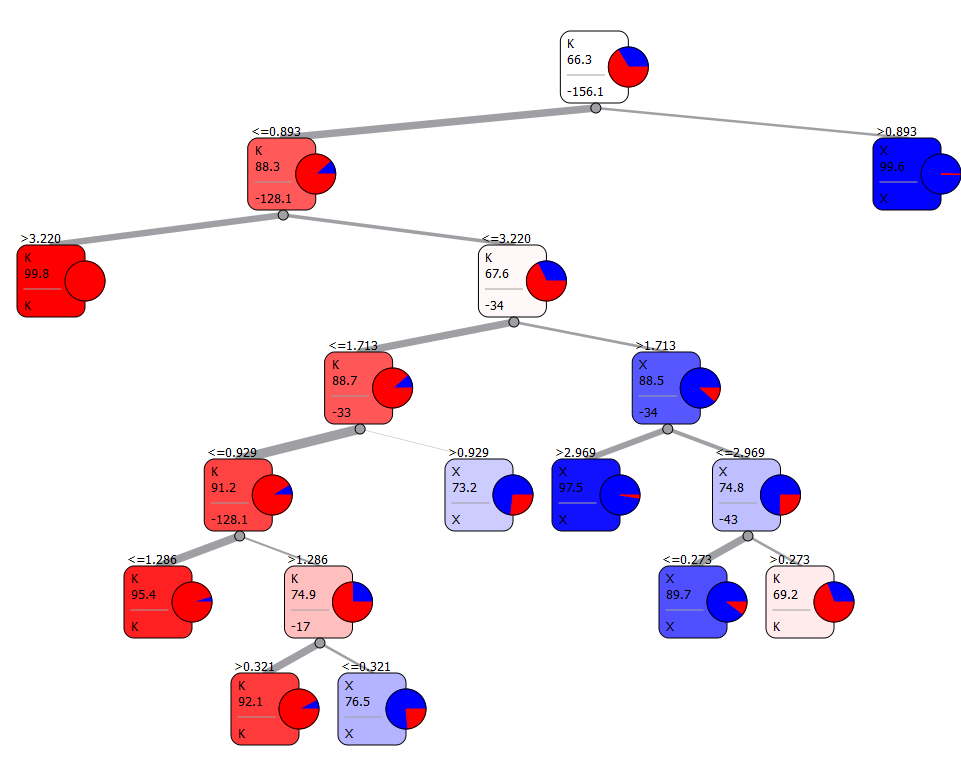


Figure 4: Decision tree to decide whether the terminal amino acid is lysine, given a charge of 1 (z=1).

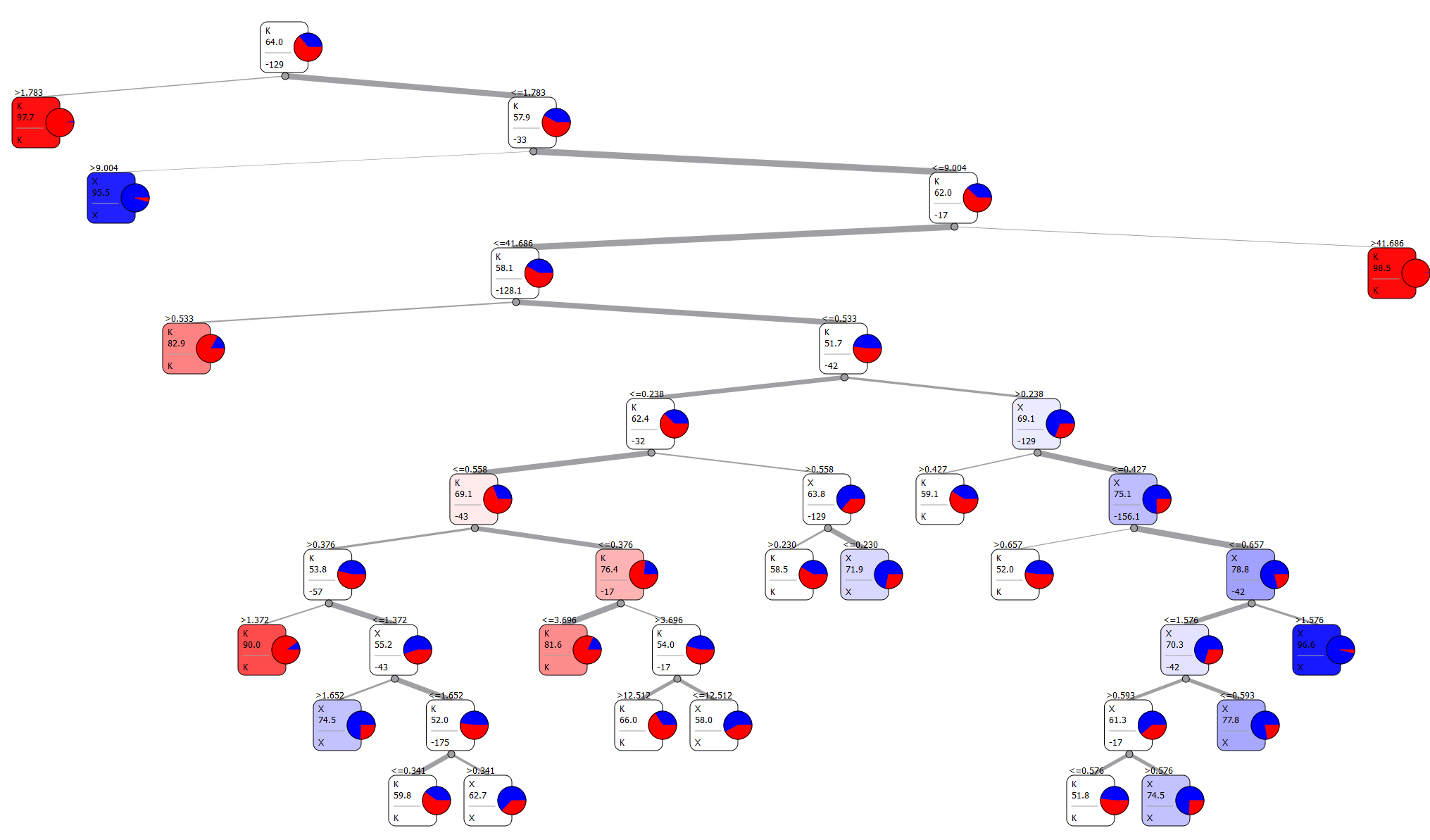


Figure 5: Decision tree to decide whether the terminal amino acid is lysine, given a charge of 2 (z=2).

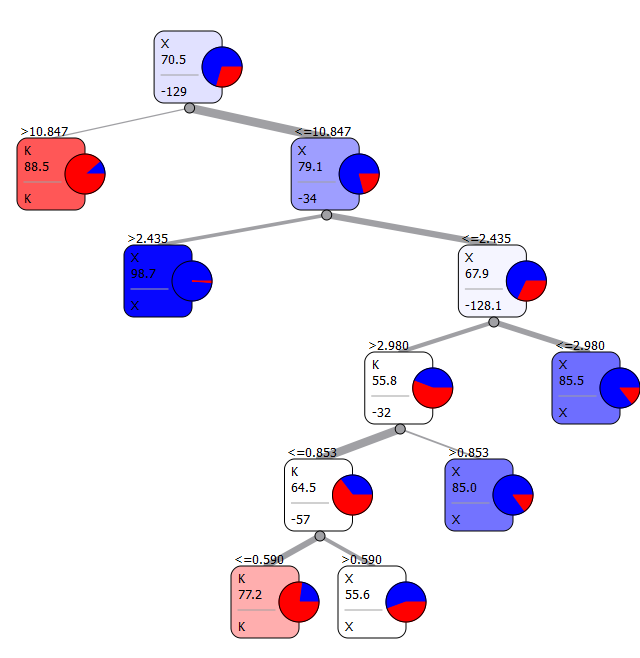


Figure 6: Decision tree to decide whether the terminal amino acid is lysine, given a charge of 3 (z=3).

# References

1. Curk T, Demsar J, Xu Q, Leban G, Petrovic U, et al. (2005) Microarray data mining with visual programming. Bioinformatics 21: 396–398. Available:http://www.ncbi.nlm.nih.gov/pubmed/15308546.