Journal of proteome research

Article

pubs.acs.org/jpr

Predicting Tryptic Cleavage from Proteomics Data Using Decision Tree Ensembles

Thomas Fannes,[§,∥] Elien Vandermarliere,[†,‡,∥] Leander Schietgat,[§] Sven Degroeve,[†,‡] Lennart Martens,*[,†,‡] and Jan Ramon[§]

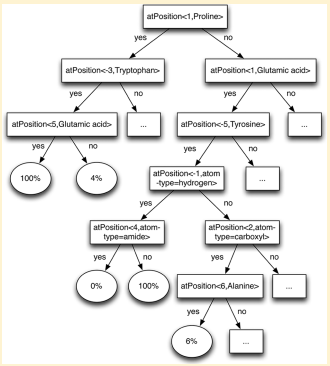
[†]Department of Medical Protein Research, VIB, B-9000 Ghent, Belgium

[‡]Department of Biochemistry, Ghent University, B-9000 Ghent, Belgium

[§]Department of Computer Science, Katholieke Universiteit Leuven, Celestijnenlaan 200A, B-3000, Leuven, Belgium

Supporting Information

ABSTRACT: Trypsin is the workhorse protease in mass spectrometry-based proteomics experiments and is used to digest proteins into more readily analyzable peptides. To identify these peptides after mass spectrometric analysis, the actual digestion has to be mimicked as faithfully as possible *in silico*. In this paper we introduce CP-DT (Cleavage Prediction with Decision Trees), an algorithm based on a decision tree ensemble that was learned on publicly available peptide identification data from the PRIDE repository. We demonstrate that CP-DT is able to accurately predict tryptic cleavage: tests on three independent data sets show that CP-DT significantly outperforms the Keil rules that are currently used to predict tryptic cleavage. Moreover, the trees generated by CP-DT can make predictions efficiently and are interpretable by domain experts.



KEYWORDS: *mass spectrometry, trypsin, PRIDE, machine learning, decision tree*

1. INTRODUCTION

In mass spectrometry-based proteomics, trypsin is the protease of choice. Trypsin, a digestive enzyme that is found in most vertebrates, cleaves the carboxy-terminal peptide bond of both Arg and Lys.[1] The resulting peptides are easily amenable to mass spectrometry, since the vast majority of their masses fall comfortably within the range of a mass spectrometer. However, cleavage by trypsin is not always reproducible nor predictable. For example, one of the most typical mistakes is the skipping of a cleavable residue (miscleavage). For trypsin, such miscleaved positions have been extensively explored, resulting in the Keil rules for miscleavage.[2] These rules predict miscleavage when (a) an Arg/Lys is followed by Pro,[3] (b) successive Lys/Arg or positive charges are close to each other, or (c) several Asp/Glu are close to the positively charged residue.[4−6] These rules have multiple limitations. For example, cleavage is rarely an all-or-nothing event, which introduces an additional level of complexity: some sites are always cleaved, other sites are never cleaved, and still other sites are cleaved with a certain probability. Furthermore, the currently known rules are unable to explain about 10% of the observed cleavages,[5] which leaves room for improvement. Finally, they were constructed on the basis of a limited amount of data.

In a shotgun proteomics experiment, which is the popular bottom-up approach to solve a proteomics question, protein identification is carried out by comparing experimental fragmentation (MS/MS) spectra, acquired from peptides obtained *via* enzymatic degradation of proteins, with theoretical spectra, derived from *in silico* digestion of sequences from protein databases.[7] The characteristics of trypsin thus play a central role in two distinct, parallel steps of this experimental setup: they not only influence the digest of the protein mixture of interest *in vitro* but also need to be mimicked *in silico* to perform the virtual digest of the protein sequence database corresponding to the sample analyzed. As such, the *in silico* workflow aims to replicate faithfully the *in vitro* workflow. This implies that, ideally, the search algorithms that perform this task have access to full knowledge on the cleavage properties of trypsin. Therefore, it is of importance that the protein cleavage pattern is predicted as accurately as possible, including any potential miscleavages. Furthermore, accurate prediction of cleavage propensity is also important in quantitative proteomics. Indeed, frequently used algorithms such as EmPAI (Exponentially modified Protein Abundance Index)[8] calculate relative quantifications of proteins by analyzing the ratio of observed peptides to observable peptides. The latter are defined as tryptic peptides with masses within the identifiable range. Accurate simulation of tryptic cleavage is particularly useful for quantification when a position can either be cleaved correctly in a variant that falls within the analytical range of the mass spectrometer or miscleaved in a variant that is too long for analysis. In this situation, the quantification relies exclusively on the observations of the correctly cleaved peptide, resulting in an

Received: February 4, 2013

   dx.doi.org/10.1021/pr4001114 | *J. Proteome Res.* XXXX, XXX, XXX−XXX

underestimation of the actual abundance of the parent protein. Correct prediction of cleavage by trypsin and other proteases is of equally great importance in targeted proteomics approaches such as Selected Reaction Monitoring (SRM). In this application, particular peptides are identified or quantified with high specificity from a complex protein mixture.[9−11] This approach, however, requires knowledge on the exact tryptic fingerprint of a protein, the cleavage positions, and if possible, the rate of cleavage have to be predicted as truthfully as possible, since only specific target peptides are selectively monitored for analysis and quantification. Even if a miscleaved version is in principle observable, it will not be a target and will therefore be missed, resulting in a wrong estimation of the actual protein concentration in the sample.

It is of note here, however, that a shotgun experiment would most likely not benefit dramatically from improved tryptic cleavage prediction, apart from a decrease in required computing time. The reason for this is that database search engines such as Mascot,[12] OMMSA,[13] and X!Tandem[14] circumvent the lack of precision by allowing one or more missed cleavages for each peptide in the *in silico* digest in order to simply cover all possibilities.[12] For the different quantification techniques, on the other hand, a large impact can be expected from knowledge on the stochastic properties of miscleavage as it allows direct correction of the estimate of protein quantities.

Recent studies already revealed some discrepancies in the Keil rules. One such study showed that the presence of cleavage between Arg/Lys and Pro is not dramatically different from the number of cleavages between Arg/Lys and Trp or Arg/Lys and Cys, both of which are generally accepted tryptic cleavage sites.[15] Another contradiction in the Keil rules is the activation, albeit slowly, of trypsin by autoproteolysis, which involves the removal of the amino-terminal peptide from its trypsinogen precursor. This autoproteolysis involves cleavage of a Lys-Ile bond that is preceded by Val-Asp-Asp-Asp-Asp, a motif that would not be cleaved according to the Keil rules.[16] These shortcomings in the Keil rules can be explained by the fact that these rules were derived several decades ago based on only a small number of experimentally confirmed cleavages. Nowadays, the mass spectrometers have a much higher analytical power and are used at very high throughput. Therefore, experiments have resulted in an overwhelming amount of data, with publicly available repositories such as PRIDE[17] increasingly capturing a substantial subset of these data. Because of the public availability of these large data sets, it has now become possible to revisit the Keil rules.

In this study, we introduce CP-DT (Cleavage Prediction with Decision Trees), which, given an amino acid sequence, is able to predict tryptic cleavage and the probability associated with the predicted cleavage. In contrast to most other tryptic prediction algorithms, CP-DT is not based on the Keil rules. Instead, it uses an ensemble of decision trees, which are well-known classifiers in machine learning.[18,19] The trees were learned on the large amount of tryptic digestion data available in PRIDE and have several advantages over the existing rules: they produce accurate predictions and can lead to knowledge that provides insight into the biology behind the predictions. They have been applied to a variety of biological problems, such as gene function prediction,[18] predicting resistance in HIV,[20] and several aspects of proteomics.[21−24] CP-DT was tested on three independent data sets, and the model was compared with the Keil rules using the area under the ROC curve (AUROC) as evaluation criterion. CP-DT achieves an AUROC of at least 83% on all test sets, outperforming the Keil rules with differences between 12% and 20%. CP-DT is freely available and can be run through a webserver at http://dtai.cs.kuleuven.be/trypsin.

2. EXPERIMENTAL SECTION

2.1. Data Sets

To develop the CP-DT algorithm, peptide data originating from experiments on human, mouse, and yeast samples retrieved from the PRIDE repository (December 2011)[17] were used as training set. In order to avoid incorrect data, all experiments were omitted that do not contain miscleavages or have a too high miscleavage ratio compared to a typical tryptic digest as determined by Foster and co-workers.[25] The miscleavage ratio of an experiment is calculated as the ratio of missed cleavages in the experiment over the total number of cleavages in that experiment. The threshold for the miscleavage ratio was set to a lower and upper limit of 0.1 and 0.4, respectively.[25] The remaining peptides were mapped onto their parent protein, and the start and end position of the peptide within the protein was determined. All identified peptides were then remapped onto their original protein. This remapping was necessary to take possible wrong annotations into account but also potential changes in UniProtKB/Swiss-Prot accession numbers due to updates in the latter repository. After these different processing steps, 481,935 peptides from human proteins originating from 866 experiments were retained. For mouse there are 20,009 peptides originating from 69 experiments, and for yeast there are 22,029 peptides from 935 experiments.

In order to evaluate CP-DT, peptide information was retrieved from three independent data resources resulting in three test sets. The first resource is an in-house data management system for MS/MS data, MS_Lims.[26] A query for shotgun experiments yielded five useful experiments containing 146,316 peptides all together. The peptides were all analyzed with an LTQ Orbitrap, either of Velos or XL type, depending on the experiment. The Mascot search engine[12] was subsequently used against the human complement of the UniProtKB/Swiss-Prot[27,28] database to identify the original proteins. The information retrieved from MS_Lims included the UniProt accession number, the start and end position of the peptide in the protein, and the peptide sequence. The full protein sequence and additional protein information was then retrieved from UniProtKB/Swiss-Prot.

The second data resource is the CPTAC (Clinical Proteomic Technology Assessment for Cancer) data set, which consists of a tryptic digest of a yeast protein mixture. In the corresponding study of the CPTAC network, the interlaboratory variation was measured by analyzing identical yeast samples in different laboratories with different instruments. In the current study, only raw data results from one laboratory were used in the evaluation (LTQ-Orbitrap@86, CPTAC experiment identification number 104).[29,30] Identification of the spectra was performed with the Mascot search engine[12] against the yeast UniProtKB/Swiss-Prot database to identify the original proteins.

The last data resource is obtained from a study performed in 2009 by the Association of Biomolecular Resource Facilities: the iPRG (The Proteome Informatics Research Group) data set. This study was performed on an *Escherichia coli* lysate and

dx.doi.org/10.1021/pr4001114 | J. Proteome Res. XXXX, XXX, XXX−XXX

detailed information on this study and the protocol can be found at http://www.abrf.org/index.cfm/group.show/ProteomicsInformaticsResearchGroup.53.htm. Here too, identification of the spectra was performed with the Mascot search engine[12] and, in this case, against the *E. coli* UniProtKB/Swiss-Prot database.

2.2. A Decision Tree Ensemble

We use decision trees to predict the probability whether a given protein sequence will be cleaved at a given cleavage position, the position of the residue just before a possible cleavage site. As trypsin cleaves exclusively the C-terminal of Lys and Arg, only positions with those two amino acids qualify as cleavage positions. Figure 1 shows an example of a decision tree. A tree

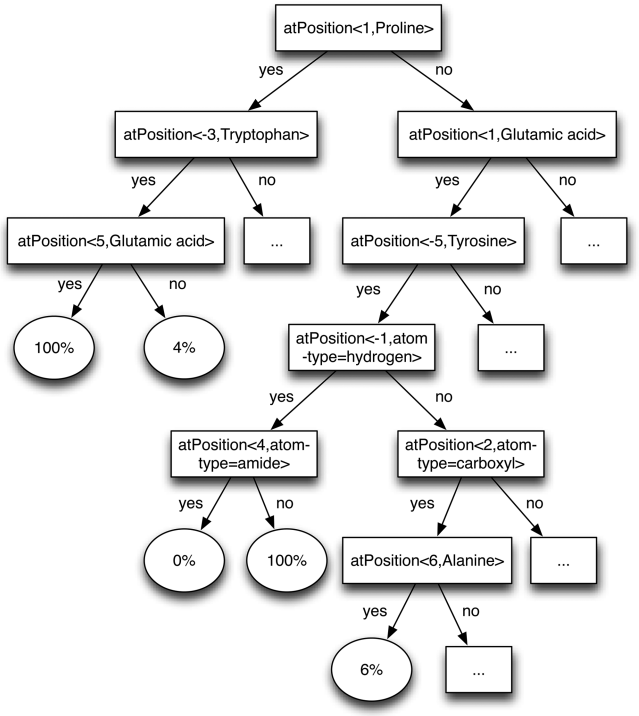


Figure 1. Example of a decision tree predicting the probability that a protein will be cleaved, using tests on the amino acid sequence window around the cleavage site.

can be regarded as a set of disjoint rules, where each rule is a conjunction of tests. A test is a boolean-valued function indicating whether a protein sequence contains a certain feature. Each internal node (depicted by a rectangle) of the tree in Figure 1 contains a test. The root node, for example, tests whether there is a Pro residue at the first position after the cleavage position. In order to predict the probability of cleavage, the cleavage position is routed down the tree according to the outcome of the tests. When a leaf node (depicted by a circle) is reached, the example is assigned the probability that is stored in the leaf. As can be seen in Figure 1, rather than returning binary values, the tree outputs a probability. If one instance has a predicted value of 0.9 and another a value of 0.6, then the cleavage is more likely to occur for the first instance than for the second. In the remainder of this section, we will explain in more detail which tests we allow in the trees, how we learn trees from labeled instances, and how we construct an ensemble of trees.

Tests. As tryptic cleavage is highly localized, tests will be restricted to a window of width 6 centered around the possible cleavage position (preliminary experiments showed that a window size of 6 is an optimal value as shown in Supplementary Table 1). The window thus contains 13 amino acids, with relative positions ranging from −6 up to 6.

We define two kinds of tests on the protein sequence and a possible cleavage position. First, the *atPosition<pos,type>* test is true if the specified *type* of amino acid occurs at position *pos*. For example, *atPosition<3,{Pro,Ala}>* returns true if there is a Pro or Ala at relative position 3 with respect to the cleavage position. Second, the *inDistance<dist,type>* test evaluates to true if there exists at least one amino acid within the specified distance from the cleavage position that is an element of the specified subset of amino acids. For example, *inDistance<3,{Pro, Ala}>* returns true if the example contains at least a Pro or an Ala between positions −3 up to +3. The position and distance parameters are of course within window range.

By having a set of the amino acids as a parameter rather than just one amino acid, the tests allow for a richer feature space, e.g., the test *atPosition<3,{Pro, Ala}>* is a logical disjunction of the tests *atPosition<3,{Pro}>* and *atPosition<3,{Ala}>*. Allowing these kinds of tests can result in smaller trees having the same expressivity.

The possible sets of amino acids that are tested are limited to groups with shared amino acid properties: name, size, charge, polarity, ring occurrence, and atom type are used to define the different subsets. A complete list of sets is defined in Supplementary Table 2, where, for example, the subset *size=tiny* defines the set of amino acids Ala, Cys, Gly, Ser, and Thr.

Top-Down Induction of Decision Trees. Classic algorithms, such as CART[31] and C4.5,[32] learn decision trees top-down. The algorithm models a decision tree based on a set of binary-labeled training instances. Each instance consists of a possible cleavage position and a label that is positive (1) if tryptic cleavage occurs at this point and negative (0) otherwise. It searches for the best acceptable test that can be put in a node. Such a test, if one can be found, induces a binary partition on the training instances: the subset for which the test returns true (the *yes* branch) and the subset for which the test returns false (the *no* branch). For both subsets in the partition, the algorithm creates a new internal node and calls itself recursively to construct a subtree. Tests receive a heuristic score, which in this case is the variance reduction. Given a set of target values of instances X and its partitions X_{yes} and X_{no} induced by test t, the variance reduction is computed as the difference in weighted variance between X and the weighted variance of each partition:

$$\sigma^2(X) - \sigma^2(X_{yes}) - \sigma^2(X_{no})$$

where the weighted variance is calculated by

$$\sigma^2(X) = \sum_{x \in X} x^2 - \frac{(\sum_{x \in X} x)^2}{|X|}$$

Minimizing variance maximizes the homogeneity of the partitions and improves predictive performance. If either the set of instances cardinality drops below a predefined threshold or no acceptable test can be found, the algorithm creates a leaf. Each leaf stores the average of the target values of its training instances, i.e., whether there occurs cleavage. For example, a value of 0.2 means that 20% of the training instances that ended up in that leaf were cleaved.

dx.doi.org/10.1021/pr4001114 | J. Proteome Res. XXXX, XXX, XXX−XXX

Ensemble of Decision Trees. Ensemble methods are learning methods that construct a set of classifiers for a given prediction task and classify new instances by combining the predictions of all classifiers. Here we consider ensembles based on random forests,[19] which is a learning technique that has primarily been used in the context of decision trees.

In our method, we construct a tree by selecting the best test from a random subset of the tests at each node. Next, multiple trees are combined into a forest. The prediction of the forest is then the average of the predictions by the individual trees.

Breiman[19] has shown that random forests can give substantial gains over predictive performance of decision tree learners. In addition, the learning time of the model rises only linearly with the size of the forest. The resulting algorithm is called CP-DT.

3. RESULTS

3.1. Experimental Setup, Algorithm Parameters and Evaluation

If the same protein is found in several experiments in the data set, we aggregated its detected peptides over all experiments. During mass spectrometry some peptides of a certain protein might not have been detected, because they were too large or too small. If for a possible cleavage site, a peptide is detected in the experiment ending at that position or starting at the next position, the site is considered as evidence of cleavage and it is labeled as a positive instance. If a peptide is found containing the position, but not at the end, the site is considered as evidence of miscleavage and we label it as a negative instance. Otherwise, the experiment is said to contain no information for the given position and we discard the site from further consideration. All labeled instances were extracted for the four available data sets, yielding the data sets given in Table 1.

Table 1. Data Set Characteristics

data set	extracted instances	species
PRIDE	681 193	human, mouse, yeast
CPTAC	23 842	yeast
iPRG	9 694	*E. coli*
MS_Lims	26 079	human

CP-DT was constructed using standard values: the ensemble consists of 100 trees, where for each node a random fraction of 10% of the number of tests was used. The minimal leaf node is 100, i.e., a node with less than 100 instances becomes a leaf node.

A model's quality is measured by the area under the ROC curve (AUROC) statistic. Based on the predicted and actual value of labeled instances in the test set, the ROC curve is calculated, and from this the AUROC can be produced. An AUROC value of 100% is a model with perfect predictive power, while a value of 50% is equivalent to a random prediction. The AUROC statistic can be used for models that output a probability, rendering it independent of a threshold function and enabling it to compare different models in a fair way.The PRIDE data set was used as training data, as it is the largest and most heterogeneous data set: it contains proteins from several species, several different settings, etc. To allow for a valid evaluation of our model, we use independent test sets. The model has not seen these data during its learning phase nor has it seen data from the same data set. We use in-house available data of tryptic digests, MS_Lims, and two publicly available data sets, CPTAC and iPRG. The two data sets were also used as control in other studies.[33,34]

3.2. Performance of CP-DT

We evaluated the Keil rule set on exactly the same data sets as mentioned above. As shown in Table 2, the Keil rule set achieves AUROC values in the range of 63% up to 73%, whereas CP-DT's values are in the range of 83% up to 90%, outperforming the Keil rules by at least 12%. A comparison of both models with respect to the three test sets can be found in Figure 2.

Table 2. Comparison of CP-DT and the Keil Rule Set Based on the AUROC Statistic

	test sets		
model	CPTAC	iPRG	MS-Lims
CP-DT	89.51%	85.76%	83.34%
Keil rule set	72.49%	73.47%	63.13%

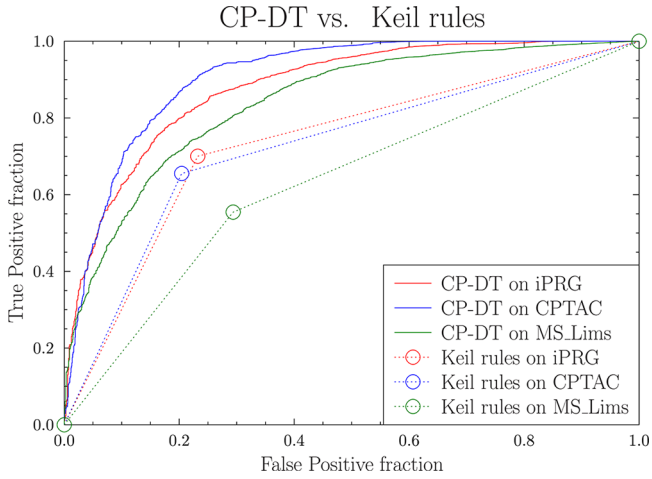


Figure 2. ROC curves of CP-DT trained on PRIDE and the Keil rules. The two different models, CP-DT model trained on the PRIDE data set on the one hand and the Keil rules on the other hand, were empirically evaluated with respect to three different data sets. It is important to note that the three test data sets used here to show the performance of CP-DT are not included in PRIDE and are therefore completely independent from the training data. The results obtained using CP-DT are shown as full lines, while the results as predicted by the Keil rules as shown as dotted lines.

Note that next to a difference in quality, there is also a difference in output expressivity: CP-DT predicts a probability, whereas the Keil rules classify an example as cleavage or miscleavage. For CP-DT, the user can decide on the ratio of true positives over false positives by choosing an appropriate threshold function.

We performed three additional experiments to gain more insight into the learned models and tryptic specificity. First, we investigated the generalizability of models learned on data originating from one specific species. Therefore, CPTAC and the yeast subset of PRIDE were used as yeast data sets, and MS_Lims and the human subset of PRIDE yielded human proteins. Models were learned on those (subsets of) data sets and evaluated on the same species data set (but not the training data set), as well as on the other species data set. The results of

dx.doi.org/10.1021/pr4001114 | J. Proteome Res. XXXX, XXX, XXX−XXX

Table 3. Comparison of AUROC When Using a Model Learned on a Single Species[a]

	test sets					
	yeast		other			
training set	CPTAC	PRIDE yeast	iPRG	MS-Lims	PRIDE human	PRIDE mouse
CPTAC	90.86%	83.62%	88.76%	82.75%	65.54%	87.74%
PRIDE yeast	89.72%	98.50%	84.57%	83.38%	72.05%	86.98%
	human		other			
	MS-Lims	PRIDE human	CPTAC	iPRG	PRIDE mouse	PRIDE yeast
MS-Lims	92.76%	61.57%	82.71%	77.50%	80.25%	80.22%
PRIDE human	81.48%	97.24%	87.24%	82.20%	86.20%	84.59%

[a]The table compares the specificity with respect to models trained on a data set from a single species and subsequently applied to an independent data set from the same species, as well as to data sets from other species. Model training was performed using yeast and human training sets.

these experiments can be found in Table 3. No significant differences in specificity across the species were found using this approach.

Second, we investigated binding site specificity for, on the one hand, different species and, on the other hand, binding site residue (Arg or Lys). The data sets were split up depending on the amino acid at the cleavage position, i.e., Arg or Lys, and different models were learned and evaluated on test sets from the same category as well as from the other category. The corresponding results are shown in Table 4. An asymmetric

Table 4. Comparison of AUROC of Models Trained with Data Sets Having Lys and Arg at the Tryptic Cleavage Site

	test sets				
	CPTAC		iPRG		MS-Lims
training sets	Lys	Arg	Lys	Arg	Lys
PRIDE Lys	90.35%	84.73%	84.95%	86.27%	83.01%
PRIDE Arg	78.25%	85.74%	65.51%	78.04%	73.38%

difference can be observed: models learned on the Lys subset have a similar quality in predicting Lys and Arg, but models learned on Arg perform significantly worse on the Lys subsets, with differences in AUROC from 12% to 20%.

Third, we want to compare the complexity (measured by the number of unique peptide sequences) of peptide databases based on our model. Therefore, CP-DT was used to predict the probability of obtaining each tryptic peptide from the digest rather than only the probability of the cleavage positions. Peptide probabilities are constructed assuming that cleavage positions are independent: the probability of a peptide is the product of the probability of cleavage at the start, the probability of cleavage at the end, and the probability that no cleavage occurred in the middle of the peptide. These predictions can then be used to reduce the complexity of the tryptic search space. From a set of proteins, a database is created that contains all peptides with a probability above a chosen threshold and a mass between 600 and 4000 Da. The number of unique peptide sequences in the database is then compared with that of the standard *in silico* tryptic digest allowing one miscleavage. Figure 3 shows the results for the human complement of the UniProt Proteomes database. The results show clearly that, for all but the very lowest peptide probability thresholds, the predictions of the CP-DT peptide set using one allowed missed cleavage overlap with those for an unlimited number of missed cleavages, indicating the ability of the algorithm to strongly reduce the complexity of the *in silico* peptide mixture regardless of the allowed number of missed cleavages.

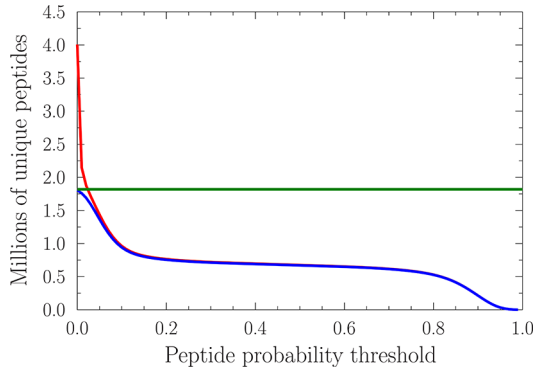


Figure 3. Comparison of the number of unique peptides in a typical tryptic digest (allowing one miscleavage) and two CP-DT predicted sets of peptides. The CP-DT peptide sets, one set with an unlimited amount of allowed missed cleavages (red line) and one set with one allowed missed cleavage (blue line), are constructed by calculating the cleavage probabilities for all peptides (see main text) and by then applying increasing peptide probability thresholds (*x*-axis). These databases are shown compared to the standard *in silico* tryptic digest database with one allowed miscleavage (green line) in terms of the number of unique peptide sequences contained in these databases. For all three databases, only peptides with a mass in between 600 and 4000 Da were retained. Overall, the CP-DT databases are much less complex than the typical *in silico* tryptic digest, and this even for the CP-DT data set that allows an unlimited number of missed cleavages. Only at extremely low probability thresholds does this data set yield a more complex *in silico* peptide mixture.

4. DISCUSSION

The introduced algorithm CP-DT uses only the primary structure of a protein to predict the probability of tryptic cleavage. There are several motivations for this choice. First, mass spectrometry-based proteomics experiments generally require full proteolysis of the proteins in the sample. This full proteolysis is typically performed on denaturated proteins, and therefore only the primary structure remains relevant. Furthermore, the amino acid sequence of a protein is most readily available for predictions, in contrast to the limited number of experimentally determined protein structures. Moreover, the generally accepted Keil rules are also based on the protein sequence alone, which allows direct and fair comparison of our predictive model and these rules.

As mentioned in Section 3.2, CP-DT outperforms the Keil rules significantly and can therefore be considered as highly

dx.doi.org/10.1021/pr4001114 | J. Proteome Res. XXXX, XXX, XXX−XXX

capable to predict tryptic cleavage. Furthermore, our results show that our model can qualitatively reduce the search space. Second, no significant difference in binding site specificity could be concluded based on the protein's species of origin. However, the specificity with respect to the residue at the binding site seems to differ: Lys-based models have a similar quality in predicting Lys and Arg instances, but Arg-based models significantly perform less on the Lys subsets. A possible reason for this effect can be that tryptic binding at an Arg is more specific than at a Lys. To allow for a more in-depth explanation of this effect, further studies are required. As the model learned on the complete PRIDE data set can discriminate between a Lys or an Arg at the binding site, it is normal that the more general model outperforms both of the models learned on the subsets.

From a machine learning point of view, we have shown that decision trees can be applied as a useful tool to learn predictive models within the field of proteomics, even when starting from highly heterogeneous public data. The technique can handle large amounts of data and extract useful signals from noisy data sets. Indeed, the model learned on PRIDE outperforms the state-of-the-art rules significantly, despite the huge variety in the underlying data sets.

■ ASSOCIATED CONTENT

Supporting Information

This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Tel: +32-92649358. Fax: +32-92649484. E-mail: lennart.martens@ugent.be.

Author Contributions

‖These authors contributed equally to this work.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors would like to thank all PRIDE submitters, the NCI CPTAC project, and the ABRF iPRG for making their data publicly available. E.V. is supported by the IWT O&O 'Kinase Switch' project, and L.M. acknowledges the support of Ghent University (Multidisciplinary Research Partnership "Bioinformatics: from nucleotides to networks") and the PRIME-XS and ProteomeXchange projects funded by the European Union seventh Framework Program under grant agreement numbers 262067 and 260558, respectively. T.F., L.S., and J.R. are supported by ERC Starting Grant 240186 "MiGraNT: Mining Graphs and Networks: a Theory-based approach". L.S. is also supported by the Research Fund KU Leuven and the FWO project "Principles of pattern set mining." This work was in part supported by the IWT SBO grant 'INSPECTOR' (120025). We thank Kurt De Grave for proofreading the text.

■ ABBREVIATIONS

CP-DT, cleavage prediction with decision trees; AUROC, area under the receiver operating characteristic curve

■ REFERENCES

(1) Rühlmann, A.; Kukla, D.; Schwager, P.; Bartels, K.; Huber, R. Structure of the complex formed by bovine trypsin and bovine pancreatic trypsin inhibitor. crystal structure determination and stereochemistry of the contact region. *J. Mol. Biol.* **1973**, *77* (3), 417−436.

(2) Keil, B. *Specificity of Proteolysis*; Springer-Verlag: Berlin, Heidelberg, NewYork, 1992.

(3) Olsen, J. V.; Ong, S.; Mann, M. Trypsin cleaves exclusively c-terminal to arginine and lysine residues. *Mol. Cell Proteomics* **2004**, *3* (6), 608−614.

(4) Siepen, J. A.; Keevil, E.; Knight, D.; Hubbard, S. J. Prediction of missed cleavage sites in tryptic peptides aids protein identification in proteomics. *J. Proteome Res.* **2007**, *6* (1), 399−408.

(5) Thiede, B.; Lamer, S.; Mattow, J.; Siejak, F.; Dimmler, C.; Rudel, T.; Jungblut, P. R. Analysis of missed cleavage sites, tryptophan oxidation and n-terminal pyroglutamylation after in-gel tryptic digestion. *Rapid Commun. Mass Spectrom.* **2000**, *14* (6), 496−502.

(6) Yen, C.; Russell, S.; Mendoza, A. M.; Meyer-Arendt, K.; Sun, S.; Cios, K. J.; Ahn, N. G.; Resing, K. A. Improving sensitivity in shotgun proteomics using a peptide-centric database with reduced complexity: protease cleavage and scx elution rules from data mining of ms/ms spectra. *Anal. Chem.* **2006**, *78* (4), 1071−1084.

(7) Frewen, B.; MacCoss, M. J. Using BiblioSpec for creating and searching tandem MS peptide libraries. *Curr. Protoc. Bioinformatics* **2007**, *Chapter 13*, Unit 13.7.

(8) Ishihama, Y.; Oda, Y.; Tabata, T.; Sato, T.; Nagasu, T.; Rappsilber, J.; Mann, M. Exponentially modified protein abundance index (empai) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol. Cell Proteomics* **2005**, *4* (9), 1265−1272.

(9) Holman, S. W.; Sims, P. F. G.; Eyers, C. E. The use of selected reaction monitoring in quantitative proteomics. *Bioanalysis* **2012**, *4* (14), 1763−1786.

(10) Lange, V.; Picotti, P.; Domon, B.; Aebersold, R. Selected reaction monitoring for quantitative proteomics: a tutorial. *Mol. Syst. Biol.* **2008**, *4*, 222.

(11) Reker, D.; Malmström, L. Bioinformatic challenges in targeted proteomics. *J. Proteome Res.* **2012**, *11* (9), 4393−4402.

(12) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, *20* (18), 3551−3567.

(13) Geer, L. Y.; Markey, S. P.; Kowalak, J. A.; Wagner, L.; Xu, M.; Maynard, D. M.; Yang, X.; Shi, W.; Bryant, S. H. Open mass spectrometry search algorithm. *J. Proteome Res.* **2004**, *3* (5), 958−964.

(14) Craig, R.; Beavis, R. C. Tandem: matching proteins with tandem mass spectra. *Bioinformatics* **2004**, *20* (9), 1466−1467.

(15) Rodriguez, J.; Gupta, N.; Smith, R. D.; Pevzner, P. A. Does trypsin cut before proline? *J. Proteome Res.* **2008**, *7* (1), 300−305.

(16) Abita, J. P.; Delaage, M.; Lazdunski, M. The mechanism of activation of trypsinogen. the role of the four n-terminal aspartyl residues. *Eur. J. Biochem.* **1969**, *8* (3), 314−324.

(17) Martens, L.; Hermjakob, H.; Jones, P.; Adamski, M.; Taylor, C.; States, D.; Gevaert, K.; Vandekerckhove, J.; Apweiler, R. Pride: the proteomics identifications database. *Proteomics* **2005**, *5* (13), 3537−3545.

(18) Schietgat, L.; Vens, C.; Struyf, J.; Blockeel, H.; Kocev, D.; Dzeroski, S. Predicting gene function using hierarchical multi-label decision tree ensembles. *BMC Bioinformatics* **2010**, *11*, 2.

(19) Breiman, L. Random forests. *Machine Learning* **2001**, *45*, 5−32.

(20) Beerenwinkel, N.; Schmidt, B.; Walter, H.; Kaiser, R.; Lengauer, T.; Hoffmann, D.; Korn, K.; Selbig, J. Diversity and complexity of hiv-1 drug resistance: a bioinformatics approach to predicting phenotype from genotype. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99* (12), 8271−8276.

(21) Qu, Y.; Adam, B. L.; Yasui, Y.; Ward, M. D.; Cazares, L. H.; Schellhammer, P. F.; Feng, Z.; Semmes, O. J.; Wright, G. L., Jr. Boosted decision tree analysis of surface-enhanced laser desorption/

ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients. *Clin. Chem.* **2002**, *48* (10), 1835−1843.
(22) Geurts, P.; Fillet, M.; De Seny, D.; Meuwis, M. A.; Malaise, M.; Merville, M. P.; Wehenkel, L. Proteomic mass spectra classification using decision tree based ensemble methods. *Bioinformatics* **2005**, *21* (14), 3138−3145.
(23) Listgarten, J.; Emili, A. Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. *Mol. Cell. Proteomics* **2005**, *4* (4), 419−434.
(24) Swaney, D. L.; McAlister, G. C.; Coon, J. J. Decision tree-driven tandem mass spectrometry for shotgun proteomics. *Nat. Methods* **2008**, *5* (11), 959−964.
(25) Foster, J. M.; Degroeve, S.; Gatto, L.; Visser, M.; Wang, R.; Griss, J.; Apweiler, R.; Martens, L. A posteriori quality control for the curation and reuse of public proteomics data. *Proteomics* **2011**, *11* (11), 2182−2194.
(26) Helsens, K.; Colaert, N.; Barsnes, H.; Muth, T.; Flikka, K.; Staes, A.; Timmerman, E.; Wortelkamp, S.; Sickmann, A.; Vandekerckhove, J.; Gevaert, K.; Martens, L. Ms_lims, a simple yet powerful open source laboratory information management system for ms-driven proteomics. *Proteomics* **2010**, *10* (6), 1261−1264.
(27) Jain, E.; Bairoch, A.; Duvaud, S.; Phan, I.; Redaschi, N.; Suzek, B. E.; Martin, M. J.; McGarvey, P.; Gasteiger, E. Infrastructure for the life sciences: design and implementation of the uniprot website. *BMC Bioinformatics* **2009**, *10*, 136.
(28) Uniprot Consortium. Ongoing and future developments at the universal protein resource. *Nucleic Acids Res.* **2011**, *39* (Database issue), D214−9.
(29) Paulovich, A. G.; Billheimer, D.; Ham, A. L.; Vega-Montoto, L.; Rudnick, P. A.; Tabb, D. L.; Wang, P.; Blackman, R. K.; Bunk, D. M.; Cardasis, H. L.; Clauser, K. R.; Kinsinger, C. R.; Schilling, B.; Tegeler, T. J.; Variyath, A. M.; Wang, M.; Whiteaker, J. R.; Zimmerman, L. J.; Fenyo, D.; Carr, S. A.; Fisher, S. J.; Gibson, B. W.; Mesri, M.; Neubert, T. A.; Regnier, F. E.; Rodriguez, H.; Spiegelman, C.; Stein, S. E.; Tempst, P.; Liebler, D. C. Interlaboratory study characterizing a yeast performance standard for benchmarking LC-MS platform performance. *Mol. Cell Proteomics* **2010**, *9* (2), 242−254.
(30) Rudnick, P. A.; Clauser, K. R.; Kilpatrick, L. E.; Tchekhovskoi, D. V.; Neta, P.; Blonder, N.; Billheimer, D. D.; Blackman, R. K.; Bunk, D. M.; Cardasis, H. L.; Ham, A. L.; Jaffe, J. D.; Kinsinger, C. R.; Mesri, M.; Neubert, T. A.; Schilling, B.; Tabb, D. L.; Tegeler, T. J.; Vega-Montoto, L.; Variyath, A. M.; Wang, M.; Wang, P.; Whiteaker, J. R.; Zimmerman, L. J.; Carr, S. A.; Fisher, S. J.; Gibson, B. W.; Paulovich, A. G.; Regnier, F. E.; Rodriguez, H.; Spiegelman, C.; Tempst, P.; Liebler, D. C.; Stein, S. E. Performance metrics for liquid chromatography-tandem mass spectrometry systems in proteomics analyses. *Mol. Cell Proteomics* **2010**, *9* (2), 225−241.
(31) Breiman, L.; Friedman, J.; Olshen, R.; Stone, C.. *Classification and Regression Trees*; Wadsworth International Group: Belmont, 1984.
(32) Quinlan, J. R. *C4.5: Programs for Machine Learning*; Morgan Kaufmann series in Machine Learning;Morgan Kaufmann Publishers: San Mateo, CA, 1993.
(33) Chang, C.; Picotti, P.; Hüttenhain, R.; Heinzelmann-Schwarz, V.; Jovanovic, M.; Aebersold, R.; Vitek, O. Protein significance analysis in selected reaction monitoring (srm) measurements. *Mol. Cell Proteomics* **2012**, *11* (4), M111.014662.
(34) Colaert, N.; Gevaert, K.; Martens, L. Ribar and xribar: methods for reproducible relative ms/ms-based label-free protein quantification. *J. Proteome Res.* **2011**, *10* (7), 3183−3189.