A Decoy-Free Approach to the Identification of Peptides

Giulia Gonnelli,‡,‡ Michiel Stock,§ Jan Verwaeren,§ Davy Maddelein,‡,‡ Bernard De Baets,§ Lennart Martens,§,†,‡ and Sven Degroeve‡,‡

INTRODUCTION

Over the past several years, mass-spectrometry-based proteomics has become the main technique for the identification and quantification of proteins in proteomics.1 In a typical workflow, experimental fragmentation spectra are acquired from peptides originating from a proteolytic digest of sample proteins, and these fragmentation spectra are subsequently matched to biological entities via specialized software tools called search engines. Several such search engines exist, such as Mascot,2 X! Tandem,3 OMSSA,4 Sequest,5 Crux,6 and Inspect.7 The identification methods used by these tools work by estimating the likelihood, in the form of a matching score, that an observed fragmentation spectrum was generated from a given peptide sequence. The output from a search engine consists of a list of possible peptide-to-spectrum matches (PSMs), associated with a score and a corresponding rank. This list is typically further processed using a statistical method to estimate the expected number of false-positives (FPs) and the false discovery rate (FDR). The FDR is typically estimated as the ratio of incorrect to correct PSMs found at a given score cutoff.8,9 In mass-spectrometry-based proteomics, so-called “target-decoy” approaches are commonly used to establish an FDR. In these approaches, the experimental spectra are also searched against artificial, “decoy” databases (often created by reversing or shuffling the sequences in the original database), and the number of hits in these databases is then used as an estimate of the number of incorrect hits obtained when searching the original database. The FDR is then estimated as the ratio of the number of decoy matches above threshold to the number of target database hits above threshold. However, this approach is not always optimal, especially in the two compelling scenarios of proteogenomics10 and metaproteomics,11 where the increased size of the search database and the high similarity between sequences make the task of separating correct from incorrect PSMs extremely challenging. Indeed, in these two scenarios, the experimental spectra are either searched against whole-genome databases that generate a large number of translated sequences that do not exist in nature (i.e., only one out of the six reading frames actually accounts for a functional protein product) or searched against a composite database compiled from many closely related genomes in the case of metaproteomics, which leads to many highly similar sequences. These factors interfere with the construction of a proper target database, which should contain only “correct” sequences, and contribute to a vast increase of potential FP hits.12 It thus becomes extremely difficult to separate correct hits from incorrect ones and to estimate the FDR in a reliable manner.13,14 Furthermore, the use of a decoy database in proteogenomics analyses of complex organisms may also be problematic for purely technical reasons: it is very hard to construct a genome-sized decoy database that does not contain any peptides observed in the original genome but that also be problematic for purely technical reasons: it is very hard to construct a genome-sized decoy database that does not contain any peptides observed in the original genome but that actually accounts for a functional protein product or searched against a composite database compiled from many closely related genomes in the case of metaproteomics, which leads to many highly similar sequences. These factors interfere with the construction of a proper target database, which should contain only “correct” sequences, and contribute to a vast increase of potential FP hits.12 It thus becomes extremely difficult to separate correct hits from incorrect ones and to estimate the FDR in a reliable manner.13,14 Furthermore, the use of a decoy database in proteogenomics analyses of complex organisms may also be problematic for purely technical reasons: it is very hard to construct a genome-sized decoy database that does not contain any peptides observed in the original genome but that
overcome the problem of searching such big, error-prone databases. All of these strategies are, however, based on either somehow reducing the size of the database or the use of initial assumptions to prevent statistical bias originating from the huge number of incorrect sequences gathered in the six-reading frame target database.

In this work we propose an alternative way to model incorrect PSMs by using lower scoring and therefore lower ranked PSMs provided by the search engine. We train a binary classifier, called Nokoi, on 47 features that describe a PSM and label correct and incorrect matches using first and lower ranks, respectively. We compare the performance of Nokoi to that of Percolator, showing very good identification performance along with short processing times. Although solutions for computing the FDR without a decoy database exist (e.g., mixture modeling the target PSM score distributions or using generating functions), decoy approaches have remained far more popular, mainly because the decoy approach does not rely on restrictive parametric assumptions.

**EXPERIMENTAL PROCEDURES**

**Training Data Retrieval**

2,229,663 PSMs were extracted from our in-house repository ms-lims. These PSMs derive from 30 distinct data sets from different samples, processed at different times on different orbitrap/ion-trap instruments. All samples used were treated with trypsin, and the precursor ion charge was limited to charges +2 and +3. For each experiment, the resulting MS/MS spectra were matched with appropriate search settings to the respective protein database using the Mascot search engine. PSMs up to rank five were collected and extracted from the resulting Mascot output files using the MascotDat file tool. Several sets of PSMs were created according to the rank assigned by Mascot: the rank1 set contains those PSMs that were ranked first by Mascot and that were scored above the Mascot significance threshold at 99% confidence, while rank2, rank3, rank4, and rank5 sets contain those PSMs ranked second to fifth, respectively. An additional decoy set, containing decoy PSMs, was created for evaluation purposes. This latter data set was obtained by searching 26,968 random subsets of MS/MS spectra from the initial 30 experiments against shuffled versions of the original target databases, yielding a grand total of 53,915 decoy hits. It must be noted that some FPs may still be present among the selected rank1 PMSs used to model correct hits as well as some false-negatives among the rank2 PSMs used to model incorrect ones. This is a well-known problem of peptide identification in MS-based proteomics, and it affects all real-world data sets. Yet even though this sort of label noise is most likely unavoidable in a general sense, this need not interfere with the creation of accurate models.

**Feature Computation**

Each PSM is represented by a 47-dimensional feature vector. Most of these features were based on the features used in Mascot Percolator and describe the PSM at the spectrum and...
Table 1. Features Used to Train the Nokoi Binary Classifier

<table>
<thead>
<tr>
<th>feature class</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>charge</td>
<td>charge of the precursor ion</td>
</tr>
<tr>
<td>log tot int</td>
<td>logarithm of the total spectrum intensity</td>
</tr>
<tr>
<td>norm highest peak</td>
<td>intensity of the highest peak in the spectrum normalized by the total spectrum intensity</td>
</tr>
<tr>
<td>pep len</td>
<td>number of amino acids in the peptide sequence</td>
</tr>
<tr>
<td>num mod</td>
<td>number of modified residues (including termini)</td>
</tr>
<tr>
<td>exp mass</td>
<td>experimental value of the precursor mass</td>
</tr>
<tr>
<td>calc mass</td>
<td>calculated value of the precursor mass</td>
</tr>
<tr>
<td>uniqueDM</td>
<td>(calc_mass - exp_mass) with isotope correction in Dalton</td>
</tr>
<tr>
<td>uniqueDMppm</td>
<td>(calc_mass - exp_mass) with isotope correction in ppm</td>
</tr>
<tr>
<td>nm pl</td>
<td>ratio of num mod to pep len</td>
</tr>
<tr>
<td>norm matched int</td>
<td>sum of the intensity of all matched fragment ions normalized by the total spectrum intensity</td>
</tr>
<tr>
<td>log matched int</td>
<td>logarithm of norm matched int</td>
</tr>
<tr>
<td>fragment ion norm int</td>
<td>normalized intensity of the matched fragment ions for the following ion series: b, y, b+, y+, b n, y n, b+n, y+n</td>
</tr>
<tr>
<td>fragment ion ratio</td>
<td>ratio of the number of matched fragment ions of the same ion series (series as above) to the total number of matched fragment ions of the spectrum</td>
</tr>
<tr>
<td>count ion series</td>
<td>number of matched fragment ions per ion series</td>
</tr>
<tr>
<td>long ion series</td>
<td>longest consecutive series of matched fragment ions for each ion series, i.e., (b1, b2, b3, b6) longest = 3 (b1, b2, b3)</td>
</tr>
<tr>
<td>medianfragDelta Da</td>
<td>median value of all matched fragment ion errors in Dalton</td>
</tr>
<tr>
<td>meanfragDelta Da</td>
<td>mean value of all matched fragment ion errors in Dalton</td>
</tr>
<tr>
<td>iqr fragDelta Da</td>
<td>interquartile range of all matched fragment ion errors in Dalton</td>
</tr>
</tbody>
</table>

along with a brief description. A complete list of all 47 features used to train the classifier is reported in Supplementary Table 1 in the SI. These 47 features were used as input features for both Nokoi and Percolator (version 2.04) during the model building and evaluation of these methods.

Model Building

Nokoi uses an L1-regularized logistic regression classifier. To train this classifier, a balanced training data set was created by randomly sampling 10000 rank1 PSMs from the initial data set as well as 10 000 rank2 PSMs. The rank1 PSMs above Mascot threshold are chosen to model correct PSMs and form the positive class, while the rank2 PSMs model the incorrect PSMs and form the negative class. When a regularized classifier is used, finding a good value for the regularization parameter is key to the performance of the classifier (as it prevents overfitting). To tune the regularization parameter (\( \lambda \)), we used a 10-fold cross-validation scheme. Within this cross-validation scheme, the following range of values for \( \lambda \) were tested: \([2^{-10}, 2^{-9}, ..., 2^5]\). Once an optimal value for \( \lambda \) was found, the model was retrained with that \( \lambda \) on the full data set. This package implements elasticnet regularization for logistic regression. The elasticnet is a generalization of L1- and L2-regularization, but with the appropriate settings an L1-regularized model can easily be obtained. To test the model that was learned, a separate test set of 20 000 PSMs, not containing the PSMs used to train the model, was sampled from the initial data set. The AUC obtained on that set was \( \sim 99\% \), showing that our model is capable of generalizing from the data. In the previous explanation, rank2 PSMs were used to form the negative class. However, the same approach was used in settings where the negative class was formed by rank3 to rank5 PSMs. Similar AUCs were obtained for these models. A plot of the largest coefficients assigned by the logistic regression to the features used by Nokoi is shown in Supplementary Figure S1 in the SI.

Evaluation Data Retrieval (CPTAC)

Publicly available data obtained from the Clinical Proteomic Tumor Analysis Consortium (CPTAC) initiative, which was supported by the NCI, were used to evaluate the models. Here, six yeast samples with spiked in Sigma 48 UPS1 standard were analyzed in triplicate by three different laboratories on LTQ-orbitrap instruments. All samples from the three laboratories together accounted for a total of \( 52 \) data sets to assess the performance of the classifier. All CPTAC data sets were searched with Mascot version 2.4.1 against a target database containing both yeast and Sigma 48 proteins. The following settings were used: acetylation of N-terminus, carbamidomethylation of cysteine, pyro-Glu formation of N-terminal glutamine, pyro-Glu formation of N-terminal glutamate, oxidation of methionine, pyro-carbamidomethylation of N-term cysteine, 10 ppm on the peptide tolerance, 0.5 Da (Da) for the fragment tolerance, 1 missed cleavage, and possible precursor charge +2, +3. Separate decoy searches were also performed against the shuffled version of the combined yeast and Sigma 48 database. To be able to use the CPTAC data as a test set, we used MascotDatfile to extract all rank1 PSMs from both target and decoy results from the Mascot searches, and the 47 features were computed for each rank1 PSM. The FDR was computed as the number of decoy hits divided by the number of target hits.

Availability

The source code and manual of all scripts making up Nokoi as well as a test data set are available online at http://genesi.ugent.be/files/costore/Nokoi_utilities.zip under the permissive Apache2 open source software license. Nokoi runs on Unix-based systems (Linux and Mac OS/X) and requires Java, Perl, and R to run.

RESULTS

Performance of the Classifier and Ability to Model PSMs with Ranks

In a usual proteomics pipeline, the experimental spectra are matched to peptide sequences through a search engine. As a result of this matching process, a list of scored PSMs is obtained. Most often, more than one peptide is found to match to a spectrum, usually with a different score. The PSM with the highest score is then chosen to be a correct match and used further in downstream analysis, while lower ranked PSMs are excluded. However, the latter PSMs could contain useful information because they represent suboptimal hits and can thus be used to model incorrect matches. Figure 2 shows the Mascot score distributions of 10 000 PSMs, sampled at random from the initial data set, belonging to rank1 scored above the Mascot significance threshold at 99% confidence, rank2, rank3, rank4, rank5, and decoys, respectively. Rank1 PSMs are clearly separated from lower ranks and decoys, while the latter are
Figure 2. Score distribution according to ranks and decoy PSMs. The distribution of highly confident Rank1 PSMs, which were above the 99\% Mascot significance threshold, is clearly well-separated from the lower ranked PSMs and decoy PSMs. Decoy PSMs and lower ranked PSMs have been assigned by Mascot with low scores and follow similar distributions. Note that the bars of the histograms corresponding to the lower ranked PSMs and the decoy PSMs are stacked for clarity.

Figure 3. Boxplots of the number of PSMs at 1\% FDR for six different models trained using different sets of PSMs: rank1 to model-correct PSMs (positive class) and the specified rank or decoy hits to model incorrect PSMs (negative class). The plot displays, for each of these six Nokoi models, the number of PSMs found at 1\% FDR across 52 CPTAC experiments. All models perform similarly, providing comparable numbers of identifications.
increasingly similar at decreasing ranks. This indicates that lower-ranked hits bear a useful similarity to decoys and can thus serve as replacements for these decoys. An additional figure of individually stacked histograms for each rank is shown in Supplementary Figure S2 in the SI. Five binary classifiers were trained on different sets of PSMs (extracted from our in-house database): Nokoi1_2, trained on rank1 (positive class) and rank2 PSMs (negative class); Nokoi1_3, on rank1 (positive class) and rank3 PSMs (negative class); Nokoi1_4, on rank1 (positive class) and rank4 PSMs (negative class); Nokoi1_5, on rank1 (positive class) and rank5 PSMs (negative class); and Nokoi1_2.3.4.5, on rank1 (positive class) and a combination of rank2, 3, 4, and 5 PSMs (negative class). The effect of subsampling is shown in Supplementary Figure S3 in the SI and shows that the model is stable. Unsurprisingly, the Nokoi1_2 model shows a wider coefficient of variation compared with the other models. This effect might be due to the higher similarity between rank1 and rank2 peptides as compared with the similarity between rank1 and lower ranked peptides. This makes the task of separating good from bad PSMs more difficult for the model trained on rank1 and rank2 peptides.12 An additional model, Nokoi1_decoy, was trained on rank1 and decoy PSMs to allow for a comparison between the decoy trained model and the other rank-based models. For each individual PSM of the test set, Nokoi outputs a score reflecting the probability of the PSM to be a correct match. The ability of each model to separate correct from incorrect matches was tested on the 52 CPTAC data sets. Figure 3 shows the performance of the six models, here computed as the number of target identifications at a fixed 1% FDR. Interestingly, all models perform similarly, indicating that ranks can be used to model correct and incorrect hits effectively and that lower ranks can be a plausible substitute for decoys. Furthermore, the Nokoi1_decoy model exhibits the lowest overall performance, suggesting that decoy PSMs may even be slightly suboptimal models of incorrect matches. This phenomenon is more pronounced when looking at the increased number of identified PSMs when using Nokoi trained on rank1 and lower ranks compared with Nokoi trained on rank1 and decoys, as reported in Supplementary Figure S4 in the SI.

Computation of the FDR Using the CPTAC Data: A Comparison of Mascot, Percolator, and Nokoi

An important aspect of shotgun proteomics is the estimation of the FDR. Indeed, when choosing a peptide identification tool, one wants to obtain the highest number of reliable peptides to be able to characterize as many proteins as possible from the original biological sample. As such, the best tools are typically the ones that are able to provide the highest number of peptides at a given FDR. State-of-the-art tools like Percolator, which works by rescoring the list of PSMs obtained as the output from a search engine, are able to significantly improve the number of correctly identified peptides. We therefore compared the number of retained peptides at a fixed 1% FDR between Nokoi, the well-known Mascot search engine, and the popular and highly performant postprocessor Percolator. As depicted in Figure 4, Nokoi identified more PSMs than Mascot yet performed slightly less well than Percolator; however, the numbers of correctly identified peptides for Nokoi and Percolator are overall quite similar. It should be noted that the overlap in identified peptide sequences for the three tools is very high, as shown in Supplementary Figure S5 in the SI. The slight advantage that Percolator has over Nokoi likely stems from the retraining of Percolator for every individual data set, while Nokoi uses a fixed prediction model that was trained only once, and this on a different (albeit very large) data set. This recurring training phase in the case of Percolator comes with a
conceivable computational cost. In contrast, Nokoi does not need to go through a training phase and simply classifies each PSM individually. Note that this also means that Nokoi can be parallelized very easily, while this is harder to achieve for Percolator. Indeed, bypassing the retraining of the algorithm coupled to the use of a simpler prediction model dramatically speeds up the classification task. An evaluation of the running time for both Nokoi and Percolator is available in the Supplementary Figures S6 and S7 in the SI.

**CONCLUSIONS**

In this work we have introduced a new approach to the concept of the target/decoy paradigm, in which the decoy PSMs are not derived from searches against dedicated decoy databases but rather are selected from the lower-ranked hits from the target database. Our tool, called Nokoi, is therefore proposed as a decoy database-free alternative that is trained on PSMs from different ranks. Our novel method is very fast and capable of quickly processing a large number of PSMs. Finally, even though Nokoi has been trained on data derived primarily from human and mouse PSMs, it has been shown to work well for the fully independent CPTAC yeast data sets. This ability to transfer the pretrained model across data sets of different origins is a key benefit of Nokoi. In summary, Nokoi demonstrates that a PSM classification model can be successfully trained using lower-ranked PSMs instead of decoy sequences and that this model shows good performance as well as broad applicability without requiring retraining.

**ASSOCIATED CONTENT**

Supporting Information

Figure S1. Boxplot of the logistic regression coefficients of the Nokoi features across 10 repetitions. Figure S2. Histograms of Mascot scores for ranked 1, 2, 3, 4, 5, and decoy PSMs. Figure S3. Boxplots of the coefficient of variation of the number of PSMs at 1% FDR measured ten times for six different Nokoi models. Figure S4. Boxplots of the increase (percentage) in the number of PSMs retained at 1% FDR between Nokoi trained on rank1 and lower ranks (Nokoi1_2.3.4.5) and Nokoi trained on rank1 and decoy PSMs. Figure S5. Percentage of peptide sequence overlap between Nokoi and Mascot/Percolator for the S2 Cptac evaluation sets. Figure S6. Running time estimation of Nokoi and Percolator on an increasing number of PSMs. Figure S7. Running time estimation of Nokoi and Percolator on the S2 CPTAC experiments. Table S1. Complete list of the 47 features used to train Nokoi and corresponding feature class as it is reported on the article. This material is available free of charge via the Internet at http://pubs.acs.org.

**AUTHOR INFORMATION**

**Corresponding Author**

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

**Notes**

The authors declare no competing financial interest.

**ACKNOWLEDGMENTS**

This work was funded by the Multidisciplinary Research Partnership “Bioinformatics: from nucleotides to networks” of Ghent University. S.D. is supported by the IWT SBO grant “INSPECTOR” (120025), and L.M. and S.D. also acknowledge support from the PRIME-XS project, grant agreement number 262067, funded by the European Union seventh Framework Program.

**ABBREVIATIONS**

PSM, peptide to spectrum matches; FDR, false discovery rate; FP, false-positives; FN, false-negatives; CPTAC, Clinical Proteomic Tumor Analysis Consortium

**REFERENCES**