Gene Expression

Annotation of tandem mass spectrometry data using stochastic neural networks in shotgun proteomics

Pavel Sulimov[1], Anastasia Voronkova[1] and Attila Kertész-Farkas [1,*]

[1] Faculty of Computer Science, School of Data Analysis and Artificial Intellingece, Moscow, 101000, Russia

* To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: The discrimination ability of score functions to separate correct from incorrect peptide-spectrum matches in database-searching-based spectrum identification are hindered by many superfluous peaks belonging to unexpected fragmentation ions or by the lacking peaks of anticipated fragmentation ions.

Results: Here, we present a new method, called BoltzMatch, to learn score functions using a particular stochastic neural networks, called restricted Boltzmann machines, in order to enhance their discrimination ability. BoltzMatch learns chemically explainable patterns among peak pairs in the spectrum data, and it can augment peaks depending on their semantic context or even reconstruct lacking peaks of expected ions during its internal scoring mechanism. As a result, BoltzMatch achieved 50% and 33% more annotations on high- and low-resolution MS2 data than XCorr at a 0.1% false discovery rate in our benchmark; conversely, XCorr yielded the same number of spectrum annotations as BoltzMatch, albeit with 4-6 times more errors. In addition, BoltzMatch alone does yield 14% more annotations than Prosit (which runs with Percolator), and BoltzMatch with Percolator yields 32% more annotations than Prosit at 0.1% FDR level in our benchmark.

Availability: BoltzMatch is freely available at: `https://github.com/kfattila/BoltzMatch`

Contact: akerteszfarkas@hse.ru

Supporting information: Supplementary materials are available at *Bioinformatics* Online.

1 Introduction

Score functions in spectrum identification (Aebersold and Mann, 2003; Nesvizhskii and Aebersold, 2004; Kertész-Farkas *et al.*, 2012; Noble and MacCoss, 2012) are hindered by (a) the presence of many unexplained peaks, which stem from the unusual fragmentation of the peptide or contaminating molecules, or (b) the lack of expected fragmentation ions, which fail to be observed in the mass spectrometer (Noble and MacCoss, 2012). Score functions attempt to mitigate the negative effects caused by these issues (a) by considering secondary fragmentation ion products (SFIP), such as the ions derived from water, carbon monoxide, or ammonia losses, in addition to primary fragmentation ions. For instance, Andromeda (Cox *et al.*, 2011) generates auxiliary peaks for water or ammonia loss products for theoretical peptides containing D, E, S, T or K, N, Q, R amino acids, respectively; while the popular XCorr function of SEQUEST (Eng *et al.*, 1994; Yates *et al.*, 1995) additionally incorporates signals from the flanking bins of the discretized spectrum vector (Eng *et al.*, 2015), SFIP, and highly charged theoretical fragmentation ion masses depending on the charge state of the precursor ion. The XCorr is formalized as

$$\mathrm{XCorr}(s,h) = E(s,h) - Z(s,h) \tag{1}$$

for a discretized experimental spectrum s and a theoretical spectrum h, where E puts a weight of 50 on the matching primary fragmentation ions, usually b and y ions, a weight of 25 on the matching flanking peaks, and a weight of 10 on the matching peaks of SFIP, and $Z(s,h)$

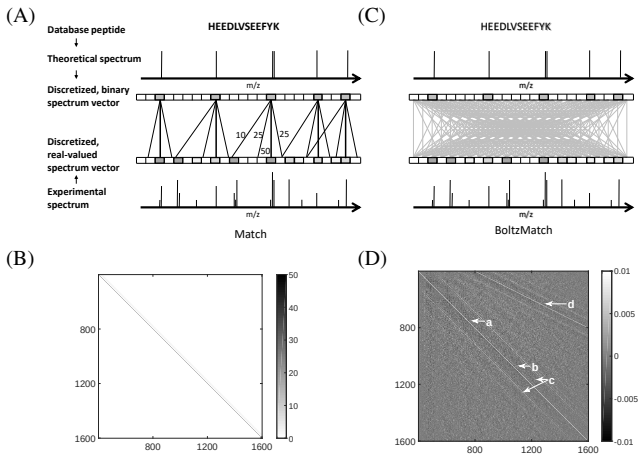

Figure 1. Graphical models of XCorr and BoltzMatch score functions along with their parameterization. (A) XCorr weights matching ions by 50, flanking peaks by 25, and losses by 10; the weight values were specified manually. (B) Heat map representation of the XCorr weights organized in a matrix. See the weights in high resolution on Supplementary Figure S1. (C) Fully connected stochastic neural network, BoltzMatch, for matching observed spectra with theoretical ones. BoltzMatch considers the association between all peak pairs and learns to weight them solely from the data. (D) Heat map representation of the weights of BoltzMatch. A positive weight $W[i, j]$ indicates an association between the peaks at positions i and j in the observed and the corresponding theoretical spectra, respectively; while a negative weight indicates an association between the peaks in the observed and the unrelated (target or decoy) theoretical spectra. The weights in the diagonal line $w[i, i]$ corresponds to matching peaks in the observed and theoretical spectra at position i indicated by arrow (a). The weights in the off-diagonal lines indicated by arrows (b) and (c) correspond to SFIP ($w[i, i+1]$, $w[i, i-1]$, $w[i, i+17]$, $w[i, i+18]$, etc.) and subsequent and antecedent primary fragmentation ions ($w[i, i+57]$, $w[i, i-57]$, $w[i, i+71]$ etc.). These distances (57, 71, etc.) relate to the masses of amino acids. The weights indicated by arrow (d) have the same interpretation as (a-c) but for double-charged peaks. See the weights in high resolution on Supplementary Figure S2.

presents a correction factor[1] (Eng *et al.*, 1994; Sulimov and Kertész-Farkas, 2019). The weights can be arranged in a weight matrix W, providing the formalization $E(s, h) = s^T W h$, where T denotes vector transposition. Figure 1A illustrates the matching of an experimental and a theoretical spectra by E, and Figure 1B shows the corresponding weight matrix W. Several new score functions and database searching tools have also been introduced, including Mascot (Perkins *et al.*, 1999), HyperScore of X!Tandem(Fenyö and Beavis, 2003), Morpheus (Wenger and Coon, 2013), and MS Amanda (Dorfer *et al.*, 2014); however, these methods are based on manually constructed score functions and have resulted in only minor improvements as compared with SEQUEST (Kim *et al.*, 2008). Recent studies focused rather on score calibration to provide a well-defined accurate semantics so that the spectrum annotations can be compared with each other (Keich and Noble, 2014; Kim and Pevzner, 2014; Kertesz-Farkas *et al.*, 2015; Keich *et al.*, 2015; Sulimov and Kertész-Farkas, 2019); however, a discussion on the discriminative power of the score functions, that is, the ability to separate correct from incorrect matches, is often neglected.

In this article, we present a novel method to learn score functions utilizing restricted Boltzmann machines (RBMs) in order to enhance the discriminative power of the score functions. RBMs are stochastic, fully connected neural networks (LeCun *et al.*, 2015) that pioneered the deep learning by being proposed as the building blocks of deep belief networks and that achieved state-of-the-art performance in various fields, such as speech recognition (Mohamed *et al.*, 2011), collaborative filtering and dimensionality reduction (Hinton and Salakhutdinov, 2006; Salakhutdinov *et al.*, 2007). In our approach, called BoltzMatch, we model the joint probability of observing an experimental s and a theoretical h spectra via RBMs, defined as

$$p(s, h) = \frac{1}{Z} \exp\{E(s, h)\}, \tag{2}$$

where the theoretical spectrum h is treated as an unobservable latent variable, an idealized version of the observed, flawed experimental spectrum s, which contains unexplainable peaks and incomplete fragmentation ion series. $E(s, h) = s^T W h$ is referred to as an energy function, and Z is a normalization factor[2], in which the parameters in W are to be learned from the observed mass spectrometry data. The log-likelihood $\log p(s, h) = E(s, z) - \log Z$ remarkably resembles the XCorr function defined in Eq. 1. On the one hand, one can roughly regard the XCorr as a log-likelihood of a manually crafted RBM, while on the other hand, one can roughly regard BoltzMatch as a generalization of XCorr in which the parameters are learned from the data. A graphical illustration of BoltzMatch and its corresponding weight matrix are shown in Figures 1C-D.

Recent machine-learning-based methods proposed for peptide identification, such as DRIP (Halloran *et al.*, 2014) or DeepNovo (Tran *et al.*, 2017), model fragmentation processes in their internal states in a temporal, sequential manner, meaning they presume that the fragmentation ions are generated one after another in time. In contrast, BoltzMatch presumes a simultaneous fragmentation process and exploits the correlations between all fragmentation ions all at once.

2 BoltzMatch

Spectrum preprocessing. Spectra were preprocessed in the same way as in the SEQUEST program but without the application of the cross-correlation penalty. Specifically, spectra were discretized along the *m/z* axis with a bin width = 1.0005079 for low-resolution MS2 settings, resulting in 2,000-dimensional spectrum vectors and a weight matrix W with a size of $2,000 \times 2,000$. For high-resolution MS2 settings spectra were discretized with a bin width = 0.05 resulting in 40,000-dimensional spectrum vectors and a weight matrix W with a size of $40,000 \times 40,000$. The discretization step was followed by the standard normalization procedure from SEQUEST (Eng *et al.*, 1994); i.e., (a) peaks around the precursor ion in a window of 1.5 Da were removed, (b) peak intensities were replaced by their square root, and (c) spectra were divided into 10 equal-length regions on the mass axis, and intensities in each segment were normalized separately. Note that the intensities were scaled to a $[0, 1]$ range in each segment. Note that the cross-correlation penalty was not applied to BoltzMatch.

Training of BoltzMatch. Training of RBMs is carried out with maximum likelihood estimation

$$\tilde{w} = \text{argmax}_w \log p_w(s) \tag{3}$$

where s denotes an experimental spectrum and p_w denotes a Gibbs distribution parameterized with w and modeled with a restricted Boltzmann machine. The weights are updated by calculating the derivatives

[1] defined as $Z(s, h) = \frac{1}{151} \sum_{\tau=-75}^{+75} E(s, h[\tau])$, where in $h[\tau]$ all vector elements of h are shifted by τ steps.

[2] defined as $Z = \sum_{s', h'} \exp\{E(s', h')\}$ for all possible vectors s', h'.

of $\log p_w(s)$ with respect to the model parameters, that is,

$$w_{i,j}^{(t+1)} = w_{i,j}^{(t)} + \frac{\partial \log p_w(s)}{\partial w_{i,j}}, \tag{4}$$

where t indicates the iteration. The derivatives lead to

$$\frac{\partial \log p_w(s)}{\partial w_{i,j}} = \sum_{h'} p_w(h' \mid s)\ s[i]\ h'[j] - \sum_{s',h'} p_w(s',h')\ s'[j]\ h'[i], \tag{5}$$

where the first summation goes over all possible binary vectors h' and the second summation goes over all possible vectors of s' and h'. The conditional probability $p_w(h \mid s)$ is defined as $p_w(h = 1 \mid s) = \prod_i \sigma(\sum_{j=1}^{n} w_{ij} v_j)$, where $\sigma(a) = (1 + \exp(-a))^{-1}$ is the sigmoid function (Fischer and Igel, 2012). The training of RBMs is notoriously hard (a) when latent variables are involved and (b) because it employs Markov chain Monte Carlo (MCMC) sampling to approximate the normalization factor Z (Hinton, 2012) to avoid the intractable enumeration of s' and h'. In order to make the training of BoltzMatch more efficient, we developed a few tricks to tackle these problems by exploiting peculiarities of the mass spectrometry data:

1. We restrict our model to only observed spectra s' and to possible theoretical spectra h' that encode real peptides. Moreover, we define $p_w(h', s') = 0$ whenever the precursor masses of these spectra are not equal up to an instrument-specific tolerance. Note that $p_w(h', s') = 0$ could lead to troubles when its logarithm is taken, thus, we just simply avoid considering such spectrum pairs in practice. Note that we consider this assumption reasonable in database-searching-base spectrum identification.
2. We assume that for every experimental spectrum there is only one theoretical peptide h that can be considered to be responsible for generating the observed spectrum s; therefore, we expect $p_w(s, h) >> 0$, while we expect $p_w(s, h') \approx 0$ for all other theoretical peptides h' within the precursor mass tolerance window. This will lead to a simplification of Eq. 5 of the following form:

$$\begin{aligned} \frac{\partial \log p_w(s)}{\partial w_{i,j}} &\approx \frac{\partial \log p_w(s,h)}{\partial w_{i,j}} \\ &= p_w(h \mid s)\ s[i]\ h[j] - \sum_{s',h'} p_w(s',h')\ s'[j]\ h'[i], \end{aligned} \tag{6}$$

   where h is the theoretical peptide responsible for generating the observed spectrum s. Unfortunately, the correct theoretical peptide is not known. Therefore, a standard database searching step is carried out to identify the (possibly) correct theoretical spectrum for each experimental spectrum with a q-value less than 0.005 prior to the training of BoltzMatch. We note that to increase the number of peptide-spectrum-matches (PSMs) with q-value less than 0.005 in our experiments, we used the Tailor score calibration method (Sulimov and Kertész-Farkas, 2019) instead of the XCorr exact p-value (XPV) method (Kim *et al.*, 2010; Howbert and Noble, 2014) because Tailor is a 2-3 magnitudes faster method than the XPV method.
3. The second summation of Eq. 5 involves the enumeration of all possible vectors h'; however, most of them do not correspond to biologically plausible vector representations of any peptides. For instance, consider a vector h in which every second bin is filled with one while all other bins are filled with zeroes; such vectors can be excluded from the enumeration. Therefore, we restrict the second summation to the candidate peptides of observed spectrum s, which leads to the following formula:

$$\begin{aligned} &\frac{\partial \log p_w(s,h)}{\partial w_{i,j}} \\ &\approx p_w(h \mid s)\ s[i]\ h[j] - \sum_{h' \in CP(s)} p_w(s,h')\ s[j]\ h'[i], \end{aligned} \tag{7}$$

   where $CP(s)$ indicates the candidate theoretical target and decoy peptides from the peptide sequence database that correspond to the experimental spectrum s. Note that in general training of RBMs, the second summation is approximated using MCMC methods; however, in our opinion, the sampling would hardly result in any biologically plausible vector that could be associated with any real peptide molecule in our case.

Observed spectrum data set can contain ubiquitous peak which appear in almost every spectrum at the same *m/z* location. For instance, the samples in the HumVar data set were prepared using TMT sixplex labeling which has an associated weight of 229.16293 Da and one can observe peaks around 230 *m/z* and 115 *m/z* in almost all experimental spectra. These peaks possibly correspond to single charge and double charged TMT labeling residues. These ubiquitous peaks do not contain useful information for spectrum identification but they interfere in generative modeling as they can correlate with all other peaks. To mitigate the effect of these ubiquitous peaks, we added a diversifying regularization (Sulimov *et al.*, 2019) in the following form:

$$DR = \sum_{s_i, s_j \in MB} h_i^T h_j, \tag{8}$$

to the learning objective defined in Eq.7, where s_i, s_j are observed spectrum pairs from a given mini-batch MB and $h_i \sim p(h_i \mid s_i) = \sigma(s^T W)$ (h_j is defined similarly), where $\sigma(a) = (1 + \exp(-a))^{-1}$ is the sigmoid function.

The training of BoltzMatch with regularization was carried out by optimizing $\log p_w(s, h)$ via maximum likelihood estimation:

$$\tilde{w} = \text{argmax}_w \left\{ \sum_{(s,h) \in D} \log \left(\frac{\exp(E_w(s,h))}{Z_s} \right) + DR \right\}, \tag{9}$$

where $Z_s = \sum_{h \in CP(s)} \exp E_w(s,h)$ and $(s,h) \in D$ denotes PSMs having q-values less than 0.005, which were obtained with standard database searching. The optimization was implemented in Python using the Pytorch toolbox, and it was run on a GTX Titan X GPU. The optimizer was a stochastic gradient descent using Nesterov momentum with a parameter of 0.9 and a learning rate of 0.001. The batch size was 128. We run the training for 15 epochs, but convergence is usually reached after 3-4 epochs. The source code is freely available in the GitHub link `www.github.com/kfattila/BoltzMatch`, and our training of all data sets can be reproduced by executing the `run_all.sh` bash script.

Evaluation of BoltzMatch in spectrum identification. Having BoltzMatch trained, PSMs were scored by $\log p(s,h) = E(s,h) - log(Z_s)$, where $Z_s = \sum_{h' \in CP(s)} E(s,h')$ and an experimental spectrum is annotated by the theoretical peptide $\tilde{h}$ that yields the highest score $\tilde{h} = \text{argmax}_{h' \in CP(s)} \log p(s,h')$. These scores are uncalibrated and proper score calibration methods can result in an increased number of spectrum annotations (Keich and Noble, 2014; Sulimov and Kertész-Farkas, 2019). We implemented the BoltzMatch scoring function in the Tide-search program from the CRUX toolkit because (a) it provides score calibration methods such as the exact p-value (XPV) method and the

Table 1. Summary of mass spectrometry data sets

Name	Instrument	#Spectra	Tolerance[1]	#Proteins[2]	#Peptides[3]	ACP[4]	MVM[5]	Modifications[6]
HumVar	LTQ Orbitrap	15,057	50 ppm/0.05 Da	91,464	3,420,673	1353.7	2	O[V],TMT6-plex[V]
Malaria	LTQ Orbitrap	12,594	50 ppm/0.05 Da	11,737	2,091,849	324.5	1	O[V],TMT6-plex[S]
iPRG	MALDI 5600	14,141	10 ppm/0.05 Da	42,450	4,283,235	185.5	1	O[V]
Aurum	MALDI 4700	9,832	2 Da/1.0005079 Da	91,464	1,591,444	7618.1	1	O[V]
HPP2A	LTQ	29,583	50 ppm/1.0005079 Da	91,464	1,591,444	443.9	1	O[V]
Yeast	LTQ	69,705	3 Da/1.0005079 Da	6,734	269,373	702.4	0	none
HeLa	Orbitrap	10,865	20 ppm/0.05 Da	193,634	4,952,900	496,3	0	none

[1]Precursor / fragment ion tolerance. No isotope error was allowed. [2]In silico enzymatic digestion was performed using the lys-C for Malaria, Trypsin/P for the HeLa, and trypsin digestion rule for the rest of the data sets. Two missed cleavages were allowed for the HeLa data set and one missed cleavages was allowed for other data sets. The minimum length of peptide was 7 and the maximum was 50 amino acids. [3] Includes modified, target and decoy peptides. [4]Average number of candidate peptides per spectrum-charge combination (ACP).The median number of candidate peptides per spectrum-charge combination can be found in the Supplementary Table S1. [5]Maximal variable modifications per peptide sequence. [6]Variable (V) and static (S) modifications, TMT-labeling (229.162932 Da) on lysine (K) and on N-terminal (Nt) modifications, oxidation (O) of methionine (+15.9949 Da). Static carbamidomethylation modification of cysteine (+57.02) was used for all data sets.

heuristic Tailor method, and (b) it can perform fast database searching for spectrum identification with all the essential parameterizations.

To calibrate the BoltzMatch score with the XPV method, let $PV_s(c)$ denote the p-value of a score c for a given spectrum s calculated with the XPV method. Then the p-value of a spectrum s and a score $c = p(s, h)$ obtained with BoltzMatch can be derived as

$$\begin{aligned} PV_s(c) &= PV_s(p(s,h)) && (10)\\ &= PV_s(\log p(s,h)) && (11)\\ &= PV_s(E(s,h) - \log Z_s) && (12)\\ &= PV_s(E(s,h)) = PV_s((s^T W)h) && (13)\\ &= PV_s(s_{BM}h), && (14) \end{aligned}$$

where Eq. 11 follows from the fact that the log function performs a monotone transformation, which does not have an impact on the p-value of any distributions, Eq. 12 follows from the fact that the normalization factor $\log Z_s$ is a spectrum-dependent constant and can be omitted, $s_{BM} = s^T W$, and $s_{BM}h$ denotes the dot product of two vectors s_{BM} and h. Therefore, the BoltzMatch scores can be calibrated with any standard XPV score calibration methods using the transformed experimental spectra s_{BM}.

Therefore, we transformed the experimental spectrum data set via $s^T W$ and exported the new spectrum data set in .ms2 file format. This was followed by a standard database search step using Tide-search program from CRUX. We made minor modifications to the Tide-search program so that it can load experimental spectrum data sets containing peak intensities with negative values and removed the application of the cross-correlation penalty procedure. This modified Tide-search/CRUX can be found at `https://github.com/kfattila/crux-toolkit`

The calibration of the BoltzMatch scores with Tailor methods is straightforward and does not require any preparation. For more details, see (Sulimov and Kertész-Farkas, 2019).

The scripts we used for training and testing along with the parameterization can be found in the GitHub repository of the BoltzMatch project.

3 Data sets and methods

3.1 Data sets

We used six, public MS2 data sets from previous publications in our experiments. Here, we give only a brief summary about the main features of the data, the detailed information about data, sample preparation, and their availability can be found in the Supplementary Notes S1. The Table 1 presents the most important parameters used in database searching for each data sets.

The **HumVar** (Human Variation) was derived from lymphoblastoid cell lines from 95 HapMap individuals, including 53 Caucasians, 33 Yorubans, 9 eastern Asians, and one Japanese (Pease *et al.*, 2013). The complete data set contains high resolution MS1 and MS2 information. In our study, we used only the `Linfeng_012511_HapMap39_3.mzML` file, which contained 15,057 experimental spectra. The **Malaria** data set is derived from a recent study of the erythrocytic cycle of the malaria parasite *Plasmodium falciparum* and obtained from (Pease *et al.*, 2013). In this study, we used only the `v07548_UofF_malaria_TMT_10.ms2` file, which contains 12,594 spectra with high resolution MS1 and MS2 information obtained with an Orbitrap spectrometer and the data was acquired from (McIlwain *et al.*, 2014). The **iPRG2012** was designed for a competition to detect modified peptides in a complex mixture by the Proteome Informatics Research Group (iPRG) at the Association of Biomolecular Resource Facilities (ABRF) (Chalkley *et al.*, 2014). The data set contains 14,141 spectra of high resolution MS1 and MS2 information generated by MALDI 5600 spectrometer from Yeast proteins. The **Aurum** data set contains 9832 singly charged spectra containing low resolution MS1 and MS2 information generated on an ABI 4700 MALDI TOF/TOF instrument from Human proteins. The **HSPP2A** data set was generated from the human protein phosphatase 2A system (Glatter *et al.*, 2009) using an LTQ mass spectrometer resulting in 29 583 spectra containing high resolution MS1 and low resolution MS2 data. The **Yeast** data set was generated from yeast (*Saccharomyces cerevisiae* strain S288C) samples using a μLC-MS/MS instrument, resulting in 69,705 spectra containing low resolution MS1 and MS2 information. Finally, the **HeLa** data set (Bekker-Jensen *et al.*, 2017) was used to compare BoltzMatch against Prosit (Gessulat *et al.*, 2019) because Prosit did not support our other data sets, it handles spectra obtained only with HCD fragmentation, and it cannot handle modifications at the time of writing. HeLa data set contained 10,865 spectra with high resolution MS1 and MS2 information from proteome of 12,250 protein-coding human genes.

3.2 Methods

Database search engines. We used the following programs: **Tide** from CRUX (McIlwain *et al.*, 2014) with XCorr score function with (a) exact p-value (XPV) calibration for low-resolution MS2 data (Howbert and Noble, 2014) and (b) Tailor calibration for high-resolution MS2 data (Sulimov and Kertész-Farkas, 2019); **Res-Ev** with XPV calibration(Lin *et al.*, 2018), **MS-GF+** (Kim and Pevzner, 2014), **X!Tandem** (Fenyö and Beavis, 2003), **OMSSA** (Geer *et al.*, 2004), **MS Amanda** (Dorfer *et al.*,

2014), **Percolator** (Käll *et al.*, 2007),**Andromeda** Cox *et al.* (2011) from MaxQuant, and **Prosit** (Gessulat *et al.*, 2019). For more details and for the parameterization of these programs, see Supplementary Note S2. The scripts we used in our experiments to run these programs can be found in the GitHub repository of the BoltzMatch project.

False discovery rate calculation. False discovery rate (FDR) was estimated using a concatenated target-decoy search (Elias and Gygi, 2007). For every target protein sequence from the FASTA file, a decoy protein sequence was generated via a protein-reverse approach, its header was appended with the label "DECOY," and the decoy protein was concatenated to the protein FASTA file. We emphasize that the database search programs were neither allowed to generate decoy peptides nor to carry out FDR estimation. The FDR estimation was carried out by ourselves using the following formula: $\tilde{FDR}(t) = \frac{\#\{Decoy>t\}+1}{\#\{Target>t\}}$ for threshold t, where $\#\{Target > t\}$ ($\#\{Decoy > t\}$) denotes the number of target (decoy) peptides identified with a PSM score larger than t. The q-values for each PSM were reported based on the FDR defined above, and the number of accepted PSMs were reported as a function of the q-values.

4 Results and Discussion

Spectrum annotation. We trained BoltzMatch using three data sets, which contained high-resolution MS2 information (HumVar, iPRG, Malaria). They contained a total of 41,792 spectra and they were discretized with a 0.05 Da bin width. We note that BoltzMatch was trained and evaluated on these data sets separately. Then we benchmarked it against search engines described above, and reported the number of accepted PSMs as a function of the q-values in Figure 4A. The results show that BoltzMatch was able to annotate 12,019 observed spectra at a 0.1% FDR, which is 49.05% more annotations compared with the standard XCorr score function when both scorings were calibrated with the Tailor method [3] (Sulimov and Kertész-Farkas, 2019). Conversely, XCorr annotated 12,019 spectra containing 5.3 times more false PSMs than BoltzMatch. Res-Ev method with XPV calibration is the current state-of-the-art scoring scheme designed specifically for high-resolution MS2 data, which was outperformed by BoltzMatch by around 17.78% more annotations at a 0.1% FDR; in contrast, Res-Ev yielded 12,019 PSMs with 4.4 times more errors than BoltzMatch.

BoltzMatch outperformed the other methods on all six data sets (HumVar, iPRG, Malaria, Aurum, HPP2a, Yeast) using a low-resolution MS2 setting, that is, all 150,912 spectra were discretized with a 1.0005079 Da bin width. BoltzMatch was trained and evaluated on these data sets separately. The results displayed in Figure 4B show that BoltzMatch was able to annotate approximately 25,288 observed spectra at a 0.1% FDR, which is 33.88% more annotations compared with the standard XCorr score function when both scorings were calibrated with the XPV method (Howbert and Noble, 2014); conversely, XCorr yielded 25,228 PSMs containing 5.3 times more errors than BoltzMatch. The detailed results obtained separately on each data sets can be seen in Supplementary Figure S3. The effects of the Diversifying Regularization and the size of the mini batch on the number of the annotations is shown and discussed in Supplementary Figure S4.

Model inspection. To reveal the reasons why BoltzMatch outperforms XCorr, we compared the transformed spectrum $s_{BM} = s^T W$ obtained with the weights of BoltzMatch and s_{XC} obtained with the application of the cross-correlation penalty of XCorr ($s_{XC} = s - \sum_{\tau=-75}^{75} s_\tau/151$, where in s_τ all vector elements are shifted by τ steps) (Eng *et al.*, 2008),

[3] Note that XPV breaks down with high-resolution MS2 data (Lin *et al.*, 2018).

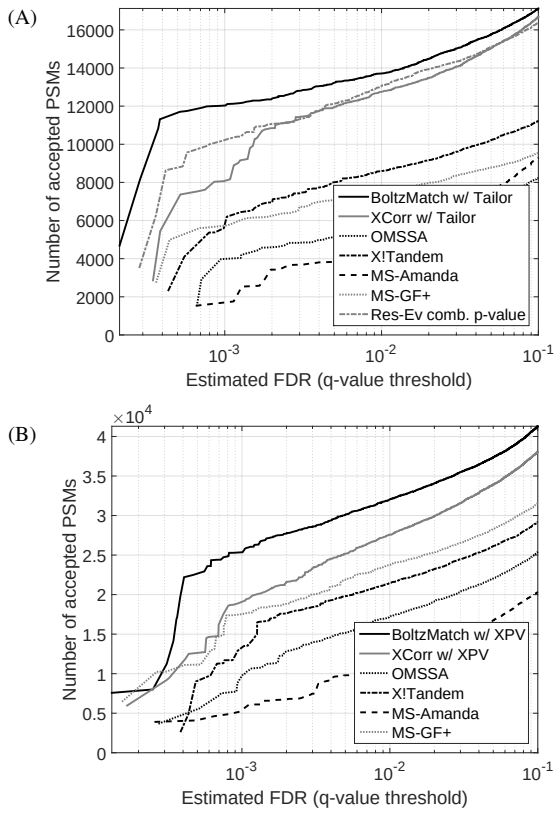

Figure 2. Spectrum annotation results with various search engines with data sets of high-resolution (panel A) and low-resolution (panel B) fragmentation information.

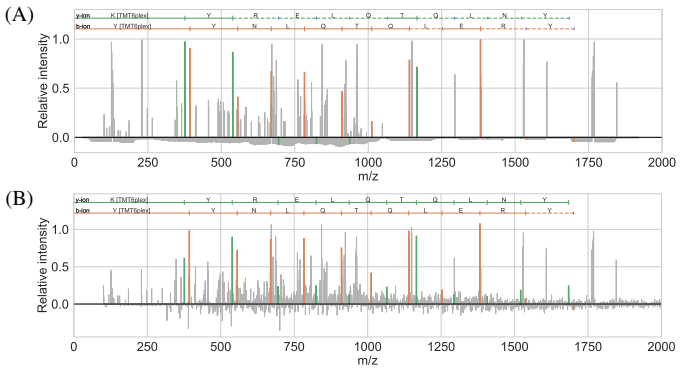

Figure 3. An observed spectrum from the Malaria data set (scan id = 7990) annotated with the database peptide YYNLQTQLERY. Peaks with positive intensity values matching to theoretical b- and y-ions are marked with green and orange colors, resp. (A) Annotation of s_{XC} obtained with XCORR with a q-value of 0.0049. (B) Annotation of s_{BM} obtained with BoltzMatch with a q-value of 0.0019.

as illustrated in Figures 3A-B, respectively, using an observed spectrum s from the Malaria data set (scan id = 7990).

On the one hand, this figure suggests that BoltzMatch normalizes the peak intensity depending on whether it can be explained by other nearby peaks, whereas XCorr diminishes peak intensities depending on the density of nearby peaks regardless of the semantic of their context. For instance, the peak corresponding to the b-ion YY (red peak near 600 *m/z* in Figure 3) was increased by around 55% with BoltzMatch but reduced by 10% with XCorr. On the other hand, BoltzMatch is able to

recover peaks corresponding to unobserved but expected fragmentation ions. For instance, the peak corresponding to the y-ion KYR (green peak near 700 *m/z* in Figure 3) was recovered from its neighboring peaks in s_{BM} with a positive intensity value, but it receives a negative intensity value (i.e. a penalty) in s_{XC} by XCorr.

Taking all annotated spectra into account, XCorr reduces the intensity of the peaks by around 5% on average regardless of whether they correspond to a fragmentation ion or not. BoltzMatch increases the intensity of the matching peaks by around 10% on average, while it keeps the intensity of the non-matching peaks intact in general. The distribution of intensity change is shown on Supplementary Figure S5. Note that BoltzMatch puts a negative weight on the peaks of the y_1 ions corresponding to the lysine (K) amino acid at the C-terminal of the peptide. This is, in fact, expected because all the theoretical peptides, both target and decoy, end with lysine at the C-terminal due to lys-C digestion; therefore, the peak of the y_1 ion uninformative for discrimination purposes.

To understand the scoring of BoltzMatch, we displayed its weights learned on Figure 1D. The visual inspection of the weights confirms that BoltzMatch is able to learn biologically and chemically plausible patterns from MS2 data, see discussion on the weights under the figure caption, and it does not require manual instrument-specific and experiment protocol-based parametrization nor does it need manual weight calibration.

Bias test. Recently, we showed that a simple machine-learning-based scoring system can easily learn to give preference to target peptides, which in turn leads to a biased FDR estimation (Danilova *et al.*, 2019). In this section, we argue on theoretical and practical grounds that BoltzMatch is not biased.

For the theoretical ground, let us consider an observed spectrum s. BoltzMatch would be biased if it assigned roughly higher scores to target peptides than to decoy peptides, that is, $\log p(s,t) \gtrapprox \log p(s,d)$ for independently sampled target t and decoy d peptides, which are unrelated to s. This would imply that the Gibbs distribution represented by a restricted Boltzmann machine has higher mass around target peptides than around decoy peptides. However, target and decoy peptides are taken from the peptide data set $t, d \in CP(S)$, and they are treated equally in the - $\sum_{h' \in CP(s)} p_w(s,h')\ s[j]\ h'[i]$ phase (called negative phase). This means that the training procedure pushes down the unnormalized probability at t and d equally, if they are unrelated to s.

For the practical ground, we took all the 15,057 top-scoring PSMs from the HumVar data set, which were obtained with BoltzMatch scoring, and selected the 5,000 worst-scoring PSMs (i.e., the bottom one-third from the ranked PSM list) for further ROC analysis. We note that the tail of a ranked PSM list should contain only incorrect spectrum annotations, which are equally likely to be matched to either target or decoy peptides in case of an unbiased scoring method. Consequently, the distribution of the target PSM scores (i.e., PSMs in which spectra matched to target peptides) and the distribution of the decoy PSM scores should be indistinguishable, which can be tested with an ROC analysis. The area under the ROC curve (AUC) obtained with our data and shown in Supplementary Figure S6 is 0.51 that has an associated p-value of 0.136 obtained with a two-sided Mann-Whitney U-test. This shows that one PSM score distribution is stochastically not greater than another one at a significance level $\alpha = 0.1$. We note that the 2.000 worst-scoring PSMs of the ranked PSM list results in an AUC of 0.49927 with an associated p-value of 0.95522.

Competition with and against other methods. Spectrum identification involves the application of various algorithms at different stages such as theoretical peptide generation methods, scoring, score calibration, post-processing algorithms. Different programs at different stages can be combined. Combining BoltzMatch scores with digestion- and charge-state-specific information using the Percolator program further improves the number of PSMs by 16% and 8% at 0.1% and 1% FDR levels (resp.) (1888 vs. 1631 at 0.1% FDR and 2071 vs. 1925 at 1% FDR) on the HeLa data set. Recently, several deep learning methods have been proposed, such as pDeep (Zhou *et al.*, 2017), Prosit (Gessulat *et al.*, 2019), DeepMass (Tiwary *et al.*, 2019), in order to improve the number of the annotations. Prosit extracts nearly 60 features for every spectrum annotation obtained with Andromeda/MaxQuant system (Cox *et al.*, 2011), including the application of the intensities of the theoretical spectra predicted with deep LSTM networks. Prosit, using the features extracted, then employs Percolator to re-score the top scoring PSMs. BoltzMatch alone annotates 14 % more PSMs than Prosit at 0.1 % FDR (1631 vs 1426), BoltzMatch+Percolator yields 32% more PSMs at 0.1% FDR (1888 vs 1426) than Prosit; however, the second yields 3% more PSMs than BoltzMatch at 1% FDR (2176 vs 2071). In fact, BoltzMatch and Prosit are not competitors; the top scoring PSMs found by BoltzMatch could be re-scored with Prosit as well; unfortunately, at the time of writing, Prosit does not support search results from search engines other than MaxQuant; thus, we could not evaluate the two methods in combination.

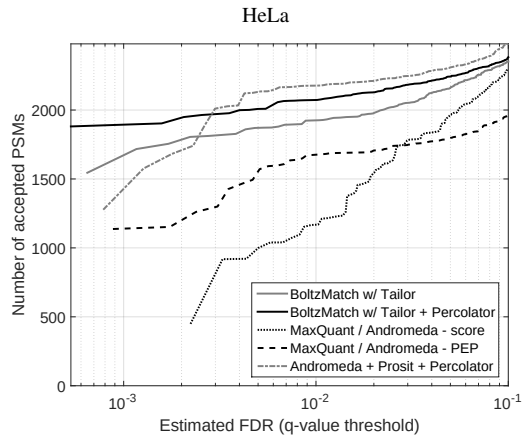

Figure 4. Search results for the HeLa data set obtained with BoltzMatch, Percolator, and Prosit. BoltzMatch outperforms Prosit at 0.1% FDR level with and without using Percolator, but it is being outperformed at 1% FDR level. Note that spectra annotated with either only CRUX or only Prosit were removed from the annotations because we were interested in comparing their scoring capability.

5 Conclusions

BoltzMatch is an interpretable, spectrum-peptide scoring method based on a fully connected stochastic neural network and is developed for database-searching-based spectrum identification in tandem mass spectrometry. BoltzMatch learns chemically explainable patterns among peak pairs of the observed and theoretical spectra, and as an outcome it may augment peaks depending on their semantic context or even reconstruct peaks of unobserved but expected fragmentation ions during its internal scoring mechanism. This information is incorporated into the scoring, which results in an increased power in discriminating between correct and incorrect spectrum annotations. Additionally, BoltzMatch does not require manual instrument-specific and experiment protocol-based parametrization such as the specification of the secondary fragmentation ions such as a ions, nor does it need manual weight calibration for the matching peaks (unlike XCorr). Unfortunately, BolztMatch is not capable of considering loss peaks conditionally for certain amino acids.

Acknowledgment

We gratefully acknowledge the support of NVIDIA Corporation for the donation of the GTX Titan X GPU used for training in this research.

References

Aebersold, R. and Mann, M. (2003). Mass spectrometry-based proteomics. *Nature*, **422**(6928), 198.

Bekker-Jensen, D. B., Kelstrup, C. D., Batth, T. S., Larsen, S. C., Haldrup, C., Bramsen, J. B., Sørensen, K. D., Høyer, S., Ørntoft, T. F., Andersen, C. L., *et al.* (2017). An optimized shotgun strategy for the rapid generation of comprehensive human proteomes. *Cell systems*, **4**(6), 587–599.

Chalkley, R. J., Bandeira, N., Chambers, M. C., Clauser, K. R., Cottrell, J. S., Deutsch, E. W., Kapp, E. A., Lam, H. H., McDonald, W. H., Neubert, T. A., *et al.* (2014). Proteome informatics research group (iprg) _2012: a study on detecting modified peptides in a complex mixture. *Molecular & Cellular Proteomics*, **13**(1), 360–371.

Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V., and Mann, M. (2011). Andromeda: a peptide search engine integrated into the maxquant environment. *Journal of Proteome Research*, **10**(4), 1794–1805.

Danilova, Y., Voronkova, A., Sulimov, P., and Kertesz-Farkas, A. (2019). Bias in false discovery rate estimation in mass-spectrometry-based peptide identification. *Journal of Proteome Research*, **18**(5), 2354–2358.

Dorfer, V., Pichler, P., Stranzl, T., Stadlmann, J., Taus, T., Winkler, S., and Mechtler, K. (2014). Ms amanda, a universal identification algorithm optimized for high accuracy tandem mass spectra. *Journal of Proteome Research*, **13**(8), 3679–3684.

Elias, J. E. and Gygi, S. P. (2007). Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature methods*, **4**(3), 207.

Eng, J. K., McCormack, A. L., and Yates, J. R. (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*, **5**(11), 976–989.

Eng, J. K., Fischer, B., Grossmann, J., and MacCoss, M. J. (2008). A fast sequest cross correlation algorithm. *Journal of Proteome Research*, **7**(10), 4598–4602.

Eng, J. K., Hoopmann, M. R., Jahan, T. A., Egertson, J. D., Noble, W. S., and MacCoss, M. J. (2015). A deeper look into comet—implementation and features. *Journal of the American Society for Mass Spectrometry*, **26**(11), 1865–1874.

Fenyö, D. and Beavis, R. C. (2003). A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Analytical chemistry*, **75**(4), 768–774.

Fischer, A. and Igel, C. (2012). An introduction to restricted boltzmann machines. In *Iberoamerican Congress on Pattern Recognition*, pages 14–36. Springer.

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

Gessulat, S., Schmidt, T., Zolg, D. P., Samaras, P., Schnatbaum, K., Zerweck, J., Knaute, T., Rechenberger, J., Delanghe, B., Huhmer, A., *et al.* (2019). Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nature methods*, **16**(6), 509.

Glatter, T., Wepf, A., Aebersold, R., and Gstaiger, M. (2009). An integrated workflow for charting the human interaction proteome: insights into the pp2a system. *Molecular systems biology*, **5**(1), 237.

Halloran, J. T., Bilmes, J. A., and Noble, W. S. (2014). Learning peptide-spectrum alignment models for tandem mass spectrometry. *Conference on Uncertainty in Artificial Intelligence*, **30**, 320.

Hinton, G. E. (2012). A practical guide to training restricted boltzmann machines. In *Neural networks: Tricks of the trade*, pages 599–619. Springer.

Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, **313**(5786), 504–507.

Howbert, J. J. and Noble, W. S. (2014). Computing exact p-values for a cross-correlation shotgun proteomics score function. *Molecular & Cellular Proteomics*, **13**(9), 2467–2479.

Käll, L., Canterbury, J. D., Weston, J., Noble, W. S., and MacCoss, M. J. (2007). Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature methods*, **4**(11), 923.

Keich, U. and Noble, W. S. (2014). On the importance of well-calibrated scores for identifying shotgun proteomics spectra. *Journal of Proteome Research*, **14**(2), 1147–1160.

Keich, U., Kertesz-Farkas, A., and Noble, W. S. (2015). Improved false discovery rate estimation procedure for shotgun proteomics. *Journal of Proteome Research*, **14**(8), 3148–3161.

Kertész-Farkas, A., Reiz, B., P Myers, M., and Pongor, S. (2012). Database searching in mass spectrometry based proteomics. *Current Bioinformatics*, **7**(2), 221–230.

Kertesz-Farkas, A., Keich, U., and Noble, W. S. (2015). Tandem mass spectrum identification via cascaded search. *Journal of proteome research*, **14**(8), 3027–3038.

Kim, S. and Pevzner, P. A. (2014). Ms-gf+ makes progress towards a universal database search tool for proteomics. *Nature communications*, **5**, 5277.

Kim, S., Gupta, N., and Pevzner, P. A. (2008). Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. *Journal of Proteome Research*, **7**(8), 3354–3363.

Kim, S., Mischerikow, N., Bandeira, N., Navarro, J. D., Wich, L., Mohammed, S., Heck, A. J., and Pevzner, P. A. (2010). The generating function of cid, etd, and cid/etd pairs of tandem mass spectra: applications to database search. *Molecular & Cellular Proteomics*, **9**(12), 2840–2852.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, **521**(7553), 436.

Lin, A., Howbert, J. J., and Noble, W. S. (2018). Combining high-resolution and exact calibration to boost statistical power: A well-calibrated score function for high-resolution ms2 data. *Journal of Proteome Research*, **17**(11), 3644–3656.

McIlwain, S., Tamura, K., Kertesz-Farkas, A., Grant, C. E., Diament, B., Frewen, B., Howbert, J. J., Hoopmann, M. R., Käll, L., Eng, J. K., *et al.* (2014). Crux: rapid open source protein tandem mass spectrometry analysis. *Journal of Proteome Research*, **13**(10), 4488–4491.

Mohamed, A.-r., Dahl, G. E., and Hinton, G. (2011). Acoustic modeling using deep belief networks. *IEEE transactions on audio, speech, and language processing*, **20**(1), 14–22.

Nesvizhskii, A. I. and Aebersold, R. (2004). Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tandem ms. *Drug discovery today*, **9**(4), 173–181.

Noble, W. S. and MacCoss, M. J. (2012). Computational and statistical analysis of protein mass spectrometry data. *PLoS computational biology*, **8**(1), e1002296.

Pease, B. N., Huttlin, E. L., Jedrychowski, M. P., Talevich, E., Harmon, J., Dillman, T., Kannan, N., Doerig, C., Chakrabarti, R., Gygi, S. P., and Chakrabarti, D. (2013). Global analysis of protein expression and phosphorylation of three stages of plasmodium falciparum intraerythrocytic development. *Journal of Proteome Research*, **12**(9), 4028–4045.

Perkins, D. N., Pappin, D. J., Creasy, D. M., and Cottrell, J. S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. *ELECTROPHORESIS: An International Journal*, **20**(18), 3551–3567.

Salakhutdinov, R., Mnih, A., and Hinton, G. (2007). Restricted boltzmann machines for collaborative filtering. In *Proceedings of the 24th International Conference on Machine learning*, pages 791–798. ACM.

Sulimov, P. and Kertész-Farkas, A. (2019). Tailor: universal, rapid, non-parametric score calibration method for database search-based peptide identification in shotgun proteomics. *bioRxiv preprint bioRxiv:10.1101/831776*.

Sulimov, P., Sukmanova, E., Chereshnev, R., and Kertész-Farkas, A. (2019). Greedy layer-wise learning of deep models using side-information. *arXiv preprint arXiv:1911.02048*.

Tiwary, S., Levy, R., Gutenbrunner, P., Soto, F. S., Palaniappan, K. K., Deming, L., Berndl, M., Brant, A., Cimermancic, P., and Cox, J. (2019). High-quality ms/ms spectrum prediction for data-dependent and data-independent acquisition data analysis. *Nature methods*, **16**(6), 519.

Tran, N. H., Zhang, X., Xin, L., Shan, B., and Li, M. (2017). De novo peptide sequencing by deep learning. *Proceedings of the National Academy of Sciences*, **114**(31), 8247–8252.

Wenger, C. D. and Coon, J. J. (2013). A proteomics search algorithm specifically designed for high-resolution tandem mass spectra. *Journal of Proteome Research*, **12**(3), 1377–1386.

Yates, J. R., Eng, J. K., McCormack, A. L., and Schieltz, D. (1995). Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Analytical chemistry*, **67**(8), 1426–1436.

Zhou, X.-X., Zeng, W.-F., Chi, H., Luo, C., Liu, C., Zhan, J., He, S.-M., and Zhang, Z. (2017). pdeep: predicting ms/ms spectra of peptides with deep learning. *Analytical chemistry*, **89**(23), 12690–12697.