On Comparison of SimTandem with State-of-the-Art Peptide Identification Tools, Efficiency of Precursor Mass Filter and Dealing with Variable Modifications

Jiří Novák[1,*], Timo Sachsenberg[2], David Hoksza[1], Tomáš Skopal[1] and Oliver Kohlbacher[2]

[1]Charles University in Prague, Faculty of Mathematics and Physics, SIRET Research Group, Malostranské nám. 25, 118 00 Prague, Czech Republic, `http://www.siret.cz`

[2]Eberhard-Karls-Universität Tübingen, Applied Bioinformatics Group, Sand 14, 72076 Tübingen, Germany, `http://abi.inf.uni-tuebingen.de`

Summary

The similarity search in theoretical mass spectra generated from protein sequence databases is a widely accepted approach for identification of peptides from query mass spectra produced by shotgun proteomics. Growing protein sequence databases and noisy query spectra demand database indexing techniques and better similarity measures for the comparison of theoretical spectra against query spectra. We employ a modification of previously proposed parameterized Hausdorff distance for comparisons of mass spectra. The new distance outperforms the original distance, the angle distance and state-of-the-art peptide identification tools OMSSA and X!Tandem in the number of identified peptides even though the q-value is only 0.001. When a precursor mass filter is used as a database indexing technique, our method outperforms OMSSA in the speed of search. When variable modifications are not searched, the search time is similar to X!Tandem. We show that the precursor mass filter is an efficient database indexing technique for high-accuracy data even though many variable modifications are being searched. We demonstrate that the number of identified peptides is bigger when variable modifications are searched separately by more search runs of a peptide identification engine. Otherwise, the false discovery rates are affected by mixing unmodified and modified spectra together resulting in a lower number of identified peptides. Our method is implemented in the freely available application SimTandem which can be used in the framework TOPP based on OpenMS.

1 Introduction

High performance liquid chromatography combined with tandem mass spectrometry (HPLC-MS/MS or shotgun proteomics) is a widely used technique for identification and quantification of proteins and peptides in complex mixtures. Mixtures obtained by a cell lysis contain thousands of proteins and a mass spectrometer produces tens of thousands of peptide mass spectra (or query spectra) which must be annotated with peptide sequences [1].

Before a mass analysis, proteins in a sample are usually enzymatically digested to peptides. After chromatographic separation, peptides are commonly subjected to an electro spray ionization

[*] To whom correspondence should be addressed. Email: novak@ksi.mff.cuni.cz

doi:10.2390/biecoll-jib-2013-228

leading to positively charged ions. After transfer into the mass spectrometer, the most intense peptide ions are collected based on their mass-to-charge ($\frac{m}{z}$) ratios and fragmented in a collision chamber. A list of $\frac{m}{z}$ ratios of fragment ions with intensities quantifying the abundance of the measured ion (i.e., a list of peaks) forms a tandem mass spectrum. The most common types of fragment ions occurring from collision induced dissociation techniques are y-ions and b-ions. Therefore, these ion types serve as main features for the annotation of spectra with peptide sequences.

The annotation of spectra with peptide sequences is often realized by means of the similarity search in databases of theoretical spectra generated from databases of known protein sequences, by the de-novo peptide sequencing, sequence-tag methods and comparison against a library of experimental spectra [2]. When the similarity search in the database of theoretical spectra is employed, protein sequences are algorithmically digested into shorter peptide sequences and theoretical peptide spectra are generated. Spectra captured by a mass spectrometer (i.e., the query set) are compared with the theoretical spectra using a pair-wise similarity function. For each query spectrum, the most similar theoretical spectrum is selected. A peptide sequence corresponding to the most similar spectrum and the query spectrum form a peptide-spectrum match (PSM). Each PSM is accompanied by the score determined by the similarity function. A natural and common similarity function for mass spectra is the cosine similarity [3]. We proposed the parameterized Hausdorff distance which is able to identify more peptides than the cosine similarity [4]. Tools based on the similarity search in databases of theoretical spectra like SEQUEST [5], MASCOT [6], OMSSA [7], X!Tandem [8] or MyriMatch [9] implement their own similarity functions.

In practice, many peptides carry additional chemical modifications which change masses of amino acids, shift $\frac{m}{z}$ ratios of fragment ions and complicate the identification of peptide sequences [10]. Modifications can be artificially added to a sample because they enable more precise analysis. They can arise during a sample preparation or during mass analysis. Post-translational modifications arise during the lifetime of a protein molecule and they give new properties to proteins, make stable conformations of proteins, regulate protein functions, etc. Protein modifications for mass spectrometry are gathered in the database UNIMOD [11] which currently contains 975 entries of known modifications.

Modifications are commonly split into two groups – *fixed* or *variable*. Fixed modifications change all amino acids of the same type in a peptide, e.g., *carbamidomethylation of cysteine*. When a fixed modification is searched, a mass of an amino acid is changed when theoretical spectra are being generated, e.g., the mass of cysteine is increased by approx. 57.02 Da. However, variable modifications do not have to change all amino acids of the same type. While processing of fixed modifications is almost for free in terms of computational complexity, processing of variable modifications is time-consuming because theoretical spectra must be generated for each combination of searched variable modifications.

Since databases of protein sequences grow rapidly in recent years, a comparison of all spectra in the query set against all theoretical spectra is time-consuming. Various database indexing techniques have been proposed to speed up the similarity search in databases of theoretical spectra. There are approaches based on the properties of metric [12] [13] and non-metric [14] [15] spaces, inverted files [16] [17], suffix trees [18], longest common prefixes and suffix arrays [19], machine learning approaches [20], support vector machines [21], neural networks [22], etc. Other approaches optimize peptide identification tools by parallelization [23], GPU processing [24], hardware acceleration [25] or by a combination of algorithmic and software engineering techniques [26] [27].

Since the search space of putative peptides can be greatly reduced by incorporating the precursor mass (i.e., the mass of a peptide ion before fragmentation), we utilize a simple database indexing technique known as the *precursor mass filter*. When the precursor mass filter is utilized, a query spectrum is not compared against all theoretical spectra generated from a database of protein sequences but only with a small subset of spectra in a precursor mass error tolerance λ. Because high-accuracy machines become easily available, the precursor mass filter is experiencing a renaissance as a database indexing technique for high-accuracy data [28] [29].

Even though different tools use different similarity functions, their performance can be compared by statistical evaluation of results [30]. A widely accepted technique is to apply a target-decoy approach. Protein sequences in a database are reversed and appended to the original database. Original sequences are marked as target sequences while reversed sequences are marked as decoy sequences. The false discovery rate can be then estimated as $FDR = \frac{\#decoy\ PSMs}{\#target\ PSMs}$. Since FDR is a property of a set of PSMs, the q-value is defined as minimum FDR threshold at which a given PSM is accepted as correct [31] [30].

2 Methods

We propose an approach for identification of peptides based on the similarity search of query spectra in a database of theoretical spectra. We describe the mass spectra distance functions, the method how we speed-up the database search using the precursor mass filter and the method how we deal with variable modifications in mass spectra. The approach is implemented in the freely available peptide identification engine SimTandem [32] which can be easily used for a batch analysis in TOPP (The OpenMS Proteomics Pipeline) [33] [34]. OpenMS is an open-source C++ library for LC-MS/MS data management and analyses. It enables a statistical evaluation of results from different peptide identification engines, thus the engines can be easily compared.

2.1 Distance Functions

When the similarity search in a database of theoretical spectra is employed for identification of peptides, a pair-wise similarity (or distance[1]) function is a crucial component of each search engine. The angle distance, the parameterized Hausdorff distance and a modification of the parameterized Hausdorff distance are defined below.

2.1.1 Angle Distance

The angle distance d_A (normalized dot product, cosine similarity) is a commonly utilized function for mass spectra comparison (Eq. 3) [3]. A representation of mass spectra as high-dimensional boolean vectors is usually used for this purpose. The range of $\frac{m}{z}$ values in a

[1] Smaller distance means bigger similarity and vice versa.

spectrum is split into subintervals. A width of a subinterval is determined by $\frac{m}{z}$ error tolerance ξ (e.g., $\xi = 0.5\,$Da). When a peak falls into a subinterval, a boolean vector contains 1 at the position corresponding to the subinterval, otherwise it contains 0 (Fig. 1).

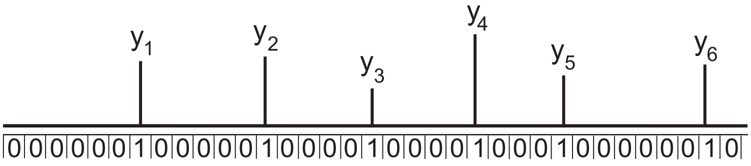

Figure 1: High-dimensional boolean representation of a theoretical spectrum containing y-ions.

Instead of storing high-dimensional sparse vectors, we use directly the vectors of $\frac{m}{z}$ values $\vec{x}$ and $\vec{y}$ (say, a low-dimensional representation of vectors). Considering the low-dimensional representation, two $\frac{m}{z}$ values between compared spectra are matched when $d_a(\vec{x_i}, \vec{y_j}) \leq \xi$. When the $\frac{m}{z}$ values are matched, the 1 is added to a sum. The $\max$ is used to prevent duplicate matches of the same $\frac{m}{z}$ value in one spectrum with more $\frac{m}{z}$ values in the other spectrum, i.e., every match of an $\frac{m}{z}$ value is counted only once. $dim(\vec{x})$ is the dimension of $\vec{x}$. Note that subintervals are not bounded as shown in Fig. 1 because the differences between $\frac{m}{z}$ values are computed.

$$d_a(\vec{x_i}, \vec{y_j}) = \begin{cases} 0, & \text{if } |\vec{x_i} - \vec{y_j}| > \xi \\ 1, & \text{else} \end{cases} \tag{1}$$

$$a(\vec{x}, \vec{y}) = \sum_{x_i \in \vec{x}} \max_{y_j \in \vec{y}} \{d_a(\vec{x_i}, \vec{y_j})\} \tag{2}$$

$$d_A(\vec{x}, \vec{y}) = \arccos\left(\frac{a(\vec{x}, \vec{y})}{\sqrt{dim(\vec{x}) dim(\vec{y})}}\right) \tag{3}$$

2.1.2 Parameterized Hausdorff Distance

The parameterized Hausdorff distance d_{HP} (Eq. 6) has been originally developed as a mass spectra distance function suitable for utilization by non-metric access methods [14] [15] [4]. For each $\frac{m}{z}$ value $\vec{x_i}$, the $\frac{m}{z}$ value $\vec{y_j}$ in the minimum distance $d_h(\vec{x_i}, \vec{y_j})$ is found (Eq. 5). Then the n^{th} root is applied on each of the minimum distances and a sum of roots is computed. When the vector $\vec{x}$ contains many irrelevant $\frac{m}{z}$ values having small differences to $\frac{m}{z}$ values in $\vec{y}$, the sum of roots generates a big distance (i.e., the similarity between $\vec{x}$ and $\vec{y}$ is poor). On the other hand, when $\vec{x}$ contains a small number of irrelevant $\frac{m}{z}$ values having big differences to $\frac{m}{z}$ values in $\vec{y}$, the sum of roots generates a small distance (i.e., the similarity between $\vec{x}$ and $\vec{y}$ is good). For $n \to \infty$, the n^{th} root converges to 1. Since numbers of $\frac{m}{z}$ values in vectors $\vec{x}$ and $\vec{y}$ may be different, the sum is divided by $dim(\vec{x})$. The whole process is repeated with vectors $\vec{x}$ and $\vec{y}$ switched and the maximum value is selected to obtain a symmetric measure. Since vectors of $\frac{m}{z}$ values are implicitly sorted, d_{HP} can be computed with linear time complexity [4].

Lets assume the following example. Let $\vec{x} = \{200, 300, 400, 500\}$ be a vector of $\frac{m}{z}$ values corresponding to a query spectrum. Let $\vec{y_1} = \{200, 300, 460, 500\}$ and $\vec{y_2} = \{210, 305, 420, 475\}$ be vectors of $\frac{m}{z}$ values corresponding to theoretical mass spectra. We can observe that $\vec{x}$ is closer to $\vec{y_1}$ in terms of mass spectra similarity. The $\frac{m}{z}$ value equals to 400 in $\vec{x}$ is likely a noise peak and the $\frac{m}{z}$ value equals to 460 is missing in $\vec{x}$. On the other hand, the spectrum $\vec{x}$ seems to be completely different from $\vec{y_2}$.

Now assume that the Euclidean distance $L_2(\vec{x},\vec{y}) = \sqrt{\sum_{i=1}^{n}(\vec{x_i} - \vec{y_i})^2}$ is used for comparison of mass spectra, then $L_2(\vec{x},\vec{y_1}) = 60$ and $L_2(\vec{x},\vec{y_2}) \doteq 33.9$. We can observe that L_2 is not suitable distance for mass spectra because $60 > 33.9$ and thus $\vec{y_2}$ is closer to $\vec{x}$ than $\vec{y_1}$. In case of d_{HP} (e.g., with $n = 2$ and $\xi = 0$), we get $d_{HP}(\vec{x},\vec{y_1}) \doteq 1.9$ and $d_{HP}(\vec{x},\vec{y_2}) \doteq 3.7$. Since $1.9 < 3.7$, the $\vec{y_1}$ is closer to $\vec{x}$ than $\vec{y_2}$ what is the desired result. For d_A ($\xi = 0$), we get $d_A(\vec{x},\vec{y_1}) \doteq 0.7$ and $d_A(\vec{x},\vec{y_2}) = \frac{\pi}{2}$. In principle, d_A and d_{HP} are similar, however, d_{HP} generates a better distribution of distances than d_A. Moreover, it has been shown that d_{HP} outperforms d_A in the number of identified peptides [15].

$$d_h(\vec{x_i},\vec{y_j}) = \begin{cases} |\vec{x_i} - \vec{y_j}|, & \text{if } |\vec{x_i} - \vec{y_j}| > \xi \\ 0, & \text{else} \end{cases} \tag{4}$$

$$h(\vec{x},\vec{y}) = \frac{\sum_{\vec{x_i}\in\vec{x}} \sqrt[n]{\min_{\vec{y_j}\in\vec{y}} \{d_h(\vec{x_i},\vec{y_j})\}}}{dim(\vec{x})} \tag{5}$$

$$d_{HP}(\vec{x},\vec{y}) = \max(h(\vec{x},\vec{y}), h(\vec{y},\vec{x})) \tag{6}$$

2.1.3 Modification of Parameterized Hausdorff Distance

We propose a modification of d_{HP} called d_{HP}^{match} (Eq. 8) to increase the number of identified peptides (Sec. 3.2). In contrast to d_{HP}, the sum of $\frac{m}{z}$ values in d_{HP}^{match} is divided by the number of matches of $\frac{m}{z}$ values in a theoretical spectrum with $\frac{m}{z}$ values in a query spectrum, i.e., $a(\vec{x},\vec{y})$ (Eq. 2). The 1 is added to $a(\vec{x},\vec{y})$ to prevent from the division by zero when $a(\vec{x},\vec{y}) = 0$.

$$h^{match}(\vec{x},\vec{y}) = \frac{\sum_{\vec{x_i}\in\vec{x}} \sqrt[n]{\min_{\vec{y_j}\in\vec{y}} \{d_h(\vec{x_i},\vec{y_j})\}}}{dim(\vec{x})(a(\vec{x},\vec{y}) + 1)} \tag{7}$$

$$d_{HP}^{match}(\vec{x},\vec{y}) = \max(h^{match}(\vec{x},\vec{y}), h^{match}(\vec{y},\vec{x})) \tag{8}$$

2.2 Precursor Mass Filter

Peptide precursor masses are known for both – theoretical and query spectra. Thus a query spectrum does not have to be compared with all theoretical spectra D generated from a database of protein sequences but only with a small subset $D_\lambda \subset D$ within a precursor mass error tolerance λ. For efficient determination of D_λ, D is sorted by precursor masses and D_λ is found by a binary search of the precursor mass of a query spectrum. Afterwards, theoretical spectra in D_λ are compared with the query spectrum using a distance function and the theoretical spectrum having the smallest distance to the query spectrum is selected to form a PSM.

2.3 Dealing with Modifications

Below, we briefly describe how we deal with variable modifications. Let m be the number of searched variable modifications and let η be the maximum number of modifications which may occur simultaneously in a peptide. A set T of all possible combinations of variable modifications is generated where each combination $t \in T$ contains up to η modifications selected from m input modifications. Because each modification can occur more than once in a peptide, the number of combinations of modifications is the sum of k-combinations with repetitions $\tau = 1 + \sum_{k=1}^{\eta} \binom{m+k-1}{k}$. The one is added to represent an unmodified peptide.

Lets assume an example where $m = 3$ and $\eta = 2$. We have three modifications α, β and γ corresponding to, e.g., *oxidation of methionine*, *dioxidation of tryptophan* and *deamidation of asparagine*. Then $\tau = 10$ combinations of modifications are generated in $T = \{\emptyset$, $\{\alpha\}$, $\{\beta\}$, $\{\gamma\}$, $\{\alpha, \alpha\}$, $\{\alpha, \beta\}$, $\{\alpha, \gamma\}$, $\{\beta, \beta\}$, $\{\beta, \gamma\}$, $\{\gamma, \gamma\}\}$. For each combination of modifications $t \in T$, the precursor mass of a query spectrum q is shifted and corresponding theoretical spectra D_λ^t in the precursor mass error tolerance λ are selected from D.

Before a theoretical spectrum from D_λ^t is compared with q, we check whether a peptide corresponding to the theoretical spectrum can contain the desired modifications. In our example, when $t = \{\alpha, \beta\}$, the peptide must contain at least one methionine and one tryptophan. When the peptide contains the desired amino acids, the theoretical spectrum is generated while masses of amino acids impacted by the modifications are shifted (i.e., the mass of methionine is shifted by α and the mass of tryptophan by β). Otherwise, the theoretical spectrum is not compared with q. When the peptide contains more than one methionine or tryptophan, all possible theoretical spectra are generated and compared with q. Finally, the theoretical spectrum having the smallest distance to q is selected from all spectra compared with q to form a PSM.

3 Results

We used HPLC-MS/MS spectra from E. coli and human. Separation of the E. coli digest was performed using an easyLC HPLC system (Proxean) with a 2h segmented gradient. Peptides eluting from the column were online injected into an LTQ-Orbitrap XL instrument (Thermo Fisher Scientific), with top 10 selection of the most abundant ions for further fragmentation. A dynamic exclusion list of 500 masses and exclusion time of 90 seconds was used to avoid repeated fragmentation of the same ions. The query set *E.coli* contained 30,358 tandem mass spectra. Human spectra were taken from 2 runs from a label-free human data set [35] – the query set *Hum48* contained 26,417 spectra and *Hum49* contained 24,537 spectra. The data sets are available on-line at [32].

The manually curated database containing 8,272 protein (332,862 peptide) sequences was used with *E.coli*. The database of 177,640 human protein (9,308,438 peptide) sequences from UniProtKB/Swiss-Prot (v. 06/2013) [36] was used with human query sets. Reversed decoy protein sequences were included in both databases. Theoretical spectra were generated with following settings – enzyme: trypsin ([KR]/P); max. missed cleavage sites: 1; length of peptide sequences: 7-50 amino acids; precursor mass of peptides: 500-5,000 Da; fragment ions types: y, b, y^{2+}; $\frac{m}{z}$ ratios of fragment ions: 200-2,000 Da; fixed modifications: carbamidomethylation of cysteine. Query spectra were processed as follows – minimum number of peaks in a spec-

Table 1: Numbers of peptides identified by different engines and search times [min:sec]. When $m = 0$, variable modifications were not searched. When $m = 5$, five variable modifications were searched. A cell having the biggest number of identified peptides among all engines is highlighted.

Query set	m	OMSSA				X!Tandem			
		q-value			Time	q-value			Time
		0.05	0.01	0.001		0.05	0.01	0.001	
E.coli	0	12,620	11,071	8,649	3:01	12,635	10,835	8,589	1:36
	5	12,841	11,248	9,009	3:56	12,807	10,942	8,510	1:43
Hum48	0	8,262	7,480	**6,646**	28:27	8,561	7,349	5,660	4:48
	5	10,806	9,598	**7,960**	30:29	11,595	9,701	7,583	5:58
Hum49	0	9,833	8,854	7,146	29:17	10,094	8,574	6,887	3:55
	5	11,742	10,477	**8,773**	31:38	12,582	10,664	8,687	5:40

Query set	m	d_A				d_{HP}			
		q-value			Time	q-value			Time
		0.05	0.01	0.001		0.05	0.01	0.001	
E.coli	0	14,024	11,204	5,340	0:37	14,173	12,228	9,551	0:41
	5	14,146	11,323	2,015	1:08	14,190	12,004	8,032	1:18
Hum48	0	7,590	4,375	879	2:58	8,666	7,172	5,333	3:40
	5	10,309	6,068	1,159	11:30	11,729	9,547	7,037	13:54
Hum49	0	9,774	6,137	1,649	3:41	10,371	8,711	7,003	3:49
	5	11,854	7,349	1,247	12:19	12,460	10,313	7,291	14:21

Query set	m	d_{HP}^{match}			
		q-value			Time
		0.05	0.01	0.001	
E.coli	0	**14,522**	**12,676**	**10,756**	0:42
	5	**14,290**	**12,437**	**9,594**	1:18
Hum48	0	**9,044**	**7,589**	6,084	3:34
	5	**12,261**	**10,268**	7,855	13:49
Hum49	0	**10,770**	**9,322**	**7,168**	3:39
	5	**13,132**	**11,110**	8,106	14:26

trum to be processed: 30; peak selection heuristic: the range of $\frac{m}{z}$ values was split by 50 Da, 5 most intense peaks were selected in each window and 50 most intense peaks were selected from the unification of the most intense peaks in the windows. $\lambda = 10\,\text{ppm}$, $\xi = 0.5\,\text{Da}$ and $n = 30$ (in d_{HP} and d_{HP}^{match}). We used SimTandem v. 1.1.65 and a machine with Windows 7 x64, Intel Core i7 2GHz, 8 GB RAM and 5400 rpm HDD.

3.1 State-of-the-Art Tools

Numbers of identified peptides for different q-values and search times were measured for freely available tools OMSSA (v. 2.1.9 Win 32) and X!Tandem (v. 2013.02.01.1). The refinement mode in X!Tandem was not used. Since X!Tandem returned some PSMs having variable modifications which were not searched, these identifications were excluded from the results. The comparison was made using OpenMS/TOPP (v. 1.10). Simple pipelines in TOPPAS were created for this purpose, e.g., *OMSSAAdapter* → *PeptideIndexer* → *FalseDiscoveryRate* → *IDFilter*, where *OMSSAAdapter* calls the OMSSA search engine, *PeptideIndexer* annotates for each search result whether it is a target or a decoy hit, *FalseDiscoveryRate* tool computes q-values and *IDFilter* selects only those PSMs with q-values less or equal a specified tolerance. The pipelines were processed without and with the support of variable modifications. When the support of variable modifications was enabled, the following five modifications were searched ($m = 5$) – *oxidation of methionine*, *deamidation of asparagine*, *acetylation of any N-term*, *pyro-glu from glutamine* and *pyro-glu from glutamic acid*.

Results are shown in Tab. 1. OMSSA identified more peptides than X!Tandem in all query sets

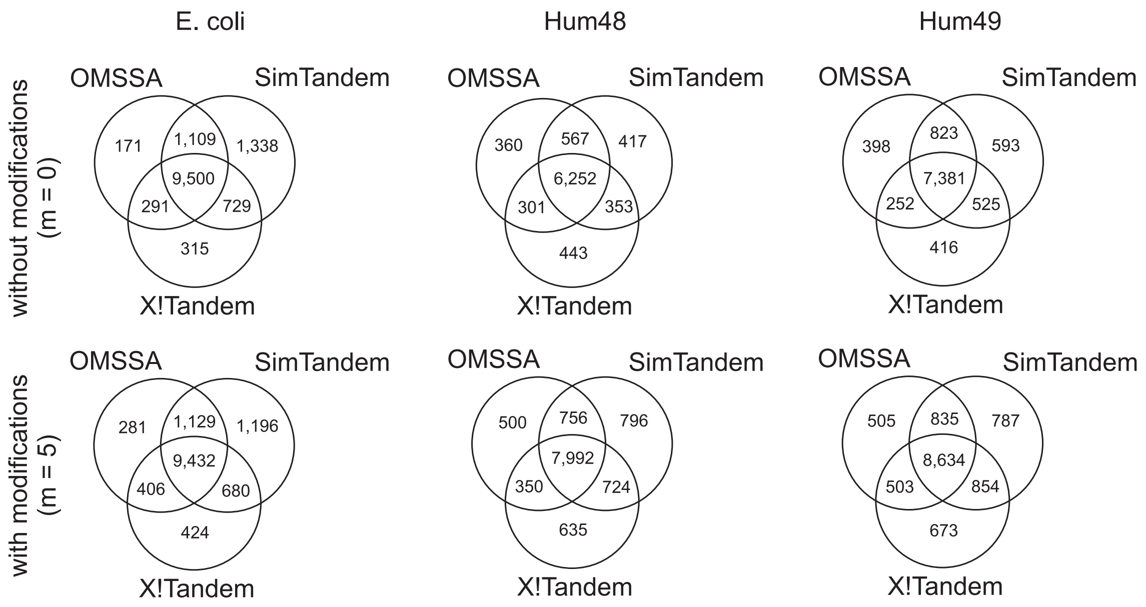

Figure 2: Numbers of peptides identified by SimTandem, OMSSA and X!Tandem (q-value $= 0.01$).

when q-value $= 0.001$. For q-value $= 0.01$, OMSSA identified more peptides than X!Tandem in four cases. However, X!Tandem identified more peptides in human query sets when variable modifications were searched. When q-value $= 0.05$, X!Tandem identified more peptides than OMSSA in all query sets except the E. coli query set when modifications were searched. X!Tandem was 1.9× faster than OMSSA on the E. coli query set when modifications were not searched and 2.3× faster when modifications were searched. On human query sets, X!Tandem was 5.9×-7.5× faster than OMSSA when modifications were not searched and 5.1×-5.6× faster when modifications were searched.

3.2 SimTandem

Numbers of peptides identified by SimTandem (i.e., by the precursor mass filter with d_A, d_{HP} or d_{HP}^{match}) and search times are shown in Tab. 1. When $m = 5$, we used $\eta = 2$. d_{HP}^{match} identified more peptide sequences than d_{HP} in all cases. The number of identified peptides was significantly smaller when d_A was used and it drastically worsened with lower q-value. d_{HP}^{match} identified more peptides than X!Tandem in all cases. OMSSA identified more peptides than d_{HP}^{match} in three cases when q-value $= 0.001$.

The overlaps of identified peptides among OMSSA, X!Tandem and SimTandem (d_{HP}^{match}) for q-value $= 0.01$ are summarized by Venn diagrams in Fig. 2. We can observe that significant numbers of peptides were identified by all three engines (from 6,252 to 9,500 peptides). The numbers of peptides identified only by SimTandem are bigger than the numbers of peptides identified only by OMSSA in all cases and the numbers of peptides identified only by X!Tandem in five cases (except *Hum48* when $m = 0$). The numbers of peptides identified only by X!Tandem are bigger than the numbers of peptides identified only by OMSSA in all cases. The numbers of peptides identified only by SimTandem and OMSSA are bigger than the numbers of peptides identified only by X!Tandem and OMSSA in all cases and bigger than the numbers of peptides identified only by SimTandem and X!Tandem in five cases (except *Hum49* when $m = 5$). The numbers of peptides identified only by SimTandem and X!Tandem are bigger than the numbers of peptides identified only by OMSSA and X!Tandem.

SimTandem (d_{HP}^{match}) was 4.3× faster than OMSSA on E. coli query set when modifications were not searched and 3× faster when modifications were searched. On human query sets, it was 8× faster than OMSSA when modifications were not searched and 2.2× faster when modifications were searched. SimTandem was 2.2× faster than X!Tandem on E. coli query set when modifications were not searched and 1.3× faster when modifications were searched. It was also 1.1×-1.4× faster than X!Tandem when modifications were not searched in human query sets. When modifications were searched in human query sets, X!Tandem was 2.3×-2.6× faster than SimTandem.

3.3 Index of the Root

Table 2: Numbers of peptides identified by d_{HP} and d_{HP}^{match} for different index n of the root. The best result in each column is highlighted.

n	q-value $= 0.001$						q-value $= 0.01$					
	d_{HP}			d_{HP}^{match}			d_{HP}			d_{HP}^{match}		
	E.coli	*Hum48*	*Hum49*	*E.coli*	*Hum48*	*Hum49*	*E.coli*	*Hum48*	*Hum49*	*E.coli*	*Hum48*	*Hum49*
1	208	23	56	2,320	200	1,009	240	23	56	6,161	1,134	1,865
2	1,411	154	346	6,578	1,747	3,355	2,502	422	504	9,959	3,528	4,613
5	5,173	1,173	2,662	10,020	4,383	6,168	7,468	2,646	3,570	12,378	6,571	8,220
10	7,554	3,691	5,230	10,478	5,425	7,169	10,555	5,132	6,345	12,602	7,314	9,060
20	9,255	4,907	6,685	10,615	5,779	7,153	11,926	6,778	8,192	12,677	7,537	9,241
30	9,551	5,333	7,003	**10,756**	6,084	7,168	12,228	7,172	8,711	12,676	7,589	**9,322**
50	10,009	5,598	7,336	10,682	6,128	7,192	12,396	7,393	8,994	12,685	7,611	9,305
100	**10,173**	**5,775**	**7,357**	10,705	**6,167**	7,053	**12,418**	**7,507**	9,117	**12,705**	**7,640**	9,294
∞	10,120	5,678	7,153	10,046	5,757	**7,223**	12,341	7,197	**9,173**	12,211	7,273	9,250

We also tested the impact of the index n of the root in d_{HP} and d_{HP}^{match} on the number of identified peptides. Variable modifications were not searched. The results are shown in Tab. 2. We can observe that the number of identified peptides is bigger with bigger n. However, when n is too big, the number of identified peptides is smaller. For both q-value $= 0.001$ and q-value $= 0.01$, the most peptides were identified in four cases when $n = 100$, in one case when $n = 30$ and in one case when $n = \infty$. In practice, the optimal n depends on the query set and should be determined empirically. Commonly, we use an empirical value $n = 30$.

3.4 Precursor Mass Filter

Since the number of comparisons of a query spectrum with theoretical spectra is crucial for the efficiency of precursor mass filter, average numbers of comparisons were measured in protein sequence databases Swiss-Prot (v. 06/2013) (human sequences only and all sequences) [36], MSDB (v. 08-Sep-2006) [37] and NCBI RefSeq (v. 55) [38]. The query set *Hum48* was used. Variable and fixed modifications were not searched. Results are shown in Tab. 3. For example, 399 theoretical spectra were compared with a query spectrum when human sequences from Swiss-Prot were used and when $\lambda = 10$ ppm. When the NCBI database was used, the number of comparisons was 60,638 for the same λ. For $\lambda = 2$ Da, the number of comparisons was significantly bigger. For example, 15,183 theoretical spectra were compared with a query

Table 3: Average numbers of comparisons of a query spectrum with theoretical spectra for different protein sequence databases and different precursor mass error tolerances λ. Numbers of protein and peptide sequences in tested protein sequence databases are also proposed (the numbers include also numbers of decoy sequences in the databases).

Database	Number of protein sequences	Number of peptide sequences	λ					
			5 ppm	10 ppm	15 ppm	0.5 Da	1 Da	2 Da
Swiss-Prot (human)	177,640	9,327,789	201	399	598	3,797	7,601	15,183
Swiss-Prot (complete)	1,080,522	52,728,460	1,063	2,106	3,157	21,404	42,923	85,714
MSDB	6,478,158	281,767,270	5,756	11,369	17,042	113,272	227,017	453,153
NCBI	34,737,538	1,533,987,691	30,606	60,638	91,004	612,225	1,227,339	2,451,235

spectrum when human sequences from Swiss-Prot were used and 2,451,235 comparisons were made when the NCBI database was used. Since the organism is usually known for a query set of spectra (e.g., E. coli or human) and the precision of modern instruments increases, the number of spectra compared with a query spectrum is small and thus the precursor filter is an efficient indexing technique for high-accuracy data.

3.5 Precursor Mass Filter and Variable Modifications

Table 4: Numbers of identified peptides, search times and total numbers of comparisons of *Hum48* with spectra generated from human protein sequences from Swiss-Prot for increasing number of searched variable modifications $m \in \langle 1, 5\rangle$ and for increasing maximum number of variable modifications in a peptide $\eta \in \langle 1, 5\rangle$.

m	Variable modifications searched			Max. number of variable modifications in a peptide η				
				1	2	3	4	5
1	oxidation of methionine	q-value	0.001	7,085	7,108	7,109	7,109	7,109
			0.01	8,669	8,685	8,680	8,680	8,680
			0.05	10,304	10,345	10,342	10,341	10,341
		Search time [min:sec]		4:29	5:07	5:24	5:59	6:03
		Num. of comparisons [millions]		14.18	14.81	14.89	14.90	14.90
2	oxidation of methionine, deamidation of asparagine	q-value	0.001	7,453	7,210	7,214	7,214	7,214
			0.01	9,243	9,263	9,244	9,245	9,245
			0.05	11,022	11,074	11,064	11,064	11,064
		Search time [min:sec]		7:02	9:01	9:32	9:59	11:27
		Num. of comparisons [millions]		20.32	25.11	26.88	27.44	27.63
3	oxidation of methionine, deamidation of asparagine, acetylation of any N-term	q-value	0.001	7,578	7,556	7,556	7,558	7,558
			0.01	9,720	9,771	9,727	9,731	9,731
			0.05	11,549	11,599	11,575	11,565	11,562
		Search time [min:sec]		8:17	13:06	16:53	18:05	20:14
		Num. of comparisons [millions]		31.36	44.88	50.57	52.45	53.05
4	oxidation of methionine, deamidation of asparagine, acetylation of any N-term, pyro-glu from glutamine	q-value	0.001	7,853	7,832	7,831	7,833	7,833
			0.01	10,140	10,242	10,172	10,176	10,176
			0.05	12,105	12,206	12,172	12,163	12,158
		Search time [min:sec]		8:42	13:40	18:46	20:40	22:41
		Num. of comparisons [millions]		32.03	46.07	51.97	53.93	54.55
5	oxidation of methionine, deamidation of asparagine, acetylation of any N-term, pyro-glu from glutamine, pyro-glu from glutamic acid	q-value	0.001	7,874	7,855	7,868	7,870	7,870
			0.01	10,186	10,268	10,209	10,213	10,213
			0.05	12,155	12,261	12,220	12,212	12,212
		Search time [min:sec]		9:33	13:49	19:26	21:39	26:37
		Num. of comparisons [millions]		33.08	47.95	54.22	56.29	56.94

We also tested the effectiveness and efficiency of SimTandem (d_{HP}^{match}) for increasing number of searched variable modifications m and for increasing maximum number of variable modifications in a peptide η. The results are presented in Tab. 4. We used the query set *Hum48* and the database of human protein sequences from Swiss-Prot. When variable modifications were not searched, the search time was 3:34 [min:sec] and the total number of comparisons of all spectra from *Hum48* against theoretical spectra was 10.58 millions of comparisons.

The search time quickly increases with bigger m and η. The reason is that theoretical spectra are generated for each combination of variable modifications. However, the total number of comparisons increases slowly because many theoretical spectra do not have to be compared with query spectra. Even though peptides corresponding to theoretical spectra have their precursor masses within λ, they do not contain amino acids affected by the searched variable modifications and thus they are not compared with query spectra (Sec. 2.3). However, the testing, whether peptides contain desired amino acids or not, causes overhead costs which increase the search time. We can reduce the search time by using $\eta \leq 2$ or $\eta \leq 3$, because the number of identified peptides does not increase significantly for bigger η.

For q-value $= 0.001$ and $m \in \langle 2, 5 \rangle$, the number of identified peptides is smaller for $\eta = 2$ than for $\eta = 1$. The same effect can be observed in all cases for q-value $= 0.01$ and q-value $= 0.05$ when $\eta = 2$ is changed to $\eta = 3$. The reason is that the spectra with variable modifications impact the distribution of target and decoy PSMs and thus they negatively impact false discovery rates and q-values [39] [40].

3.6 FDRs of Spectra with Variable Modifications

Table 5: Numbers of PSMs having variable modifications and search times [min:sec] in two cases – when variable modifications are searched separately in five search runs ($m = 1$) and when variable modifications are searched together in a search run ($m = 5$).

Variable modifications searched	*E.coli*				*Hum48*				*Hum49*			
	q-value			Time	q-value			Time	q-value			Time
	0.05	0.01	0.001		0.05	0.01	0.001		0.05	0.01	0.001	
oxidation of methionine	823	582	421	0:44	1,327	1,114	914	4:29	1,684	1,454	1,185	4:37
deamidation of asparagine	394	177	116	0:44	911	642	441	5:09	659	463	310	5:02
acetylation of any N-term	251	56	12	0:47	774	512	360	6:17	346	144	98	6:20
pyro-glu from glutamine	139	110	92	0:42	502	428	284	4:02	195	169	134	4:11
pyro-glu from glutamic acid	41	12	4	0:43	83	50	27	4:30	101	67	31	4:32
Total	1,648	937	645	3:40	3,597	2,746	2,026	24:27	2,985	2,297	1,758	24:42
oxidation of methionine, deamidation of asparagine, acetylation of any N-term, pyro-glu from glutamine, pyro-glu from glutamic acid	1,207	820	558	0:59	3,469	2,706	1,985	9:33	2,742	2,211	1,598	9:39

Since many searched variable modifications may impact the q-values, we have compared the numbers of PSMs having variable modifications in two cases. First, we searched $m = 5$ modifications together in a search run of the peptide identification engine. Second, we run the engine for each modification separately (i.e., we performed five searches when $m = 1$). We used d_{HP}^{match} and $\eta = 1$. The results are summarized in Tab. 5. The total number of identified PSMs from independent search runs is bigger than the number of PSMs identified in the search run where all the modifications are searched together. Even though the searching for each modification separately is time-consuming (2.6×-3.7× slower) because the search engine must be used many times, the approach might be interesting for practical usage because of bigger number of identified peptide sequences. The advantage of this approach has been also emphasized in [39].

4 Conclusion

We have proposed a method for identification of peptides from tandem mass spectra based on the similarity search in databases of theoretical spectra generated from databases of known protein sequences. Our method employs a modification of parameterized Hausdorff distance which outperforms the original distance and the angle distance in the number of identified peptides. Moreover, it outperforms state-of-the-art peptide identification tools OMSSA and X!Tandem. When the precursor mass filter is utilized as a database indexing technique, our method is faster than OMSSA. When variable modifications are not being searched, its search time is similar to the search time of X!Tandem. We have studied the efficiency of precursor mass filter considering different protein sequence databases and different precursor mass error tolerances. Since the accuracy of modern instruments increases in recent years, the precursor mass filter is an efficient database indexing technique for high-accuracy data.

We analyzed the numbers of identified peptides and search times when variable modifications were searched. Generally, when the maximum number of variable modifications in a peptide is set up to 2 or 3, we can reduce the search time even though many variable modifications are being searched. However, the number of identified peptides is smaller with bigger number of searched variable modifications because the computation of false discovery rates is affected by mixing of modified and unmodified spectra together. Thus it seems to be advantageous to run the peptide identification engine for each variable modification or a small set of variable modifications separately. Our method is implemented in the freely available peptide identification engine SimTandem which can be used for a batch analysis in TOPP based on OpenMS. Moreover, our results can be easily reproduced by TOPP.

Acknowledgements

This work was supported in part by Czech Science Foundation (GAČR) project Nr. 202/11/0968 and by the grant SVV-2013-267312.

References

[1] I. Eidhammer, K. Flikka, L. Martens and S.-O. Mikalsen. *Computational Methods for Mass Spectrometry Proteomics*. John Wiley & Sons, England, 2007.

[2] A. I. Nesvizhskii. A Survey of Computational Methods and Error Rate Estimation Procedures for Peptide and Protein Identification in Shotgun Proteomics. *Journal of Proteomics*, 73(11):2092–2123, 2010.

[3] J. Liu, A. Bell, J. Bergeron, C. Yanofsky, B. Carrillo, C. Beaudrie and R. Kearney. Methods for Peptide Identification by Spectral Comparison. *Proteome Science*, 5(1):3, 2007.

[4] J. Novák and D. Hoksza. Parametrised Hausdorff Distance as a Non-Metric Similarity Model for Tandem Mass Spectrometry. In *CEUR Proc. DATESO*, pages 1–12. 2010.

[5] J. Eng, A. L. McCormack and J. R. Yates III. 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, 1994.

[6] D. N. Perkins, D. J. C. Pappin, D. M. Creasy and J. S. Cottrell. Probability-based Protein Identification by Searching Sequence Databases using Mass Spectrometry Data. *Electrophoresis*, 20(18):3551–3567, 1999.

[7] L. Y. Geer, S. P. Markey, J. A. Kowalak, L. Wagner, M. Xu, D. M. Maynard, X. Yang, W. Shi and S. H. Bryant. Open Mass Spectrometry Search Algorithm. *Journal of Proteome Research*, 3(5):958–964, 2004.

[8] R. Craig and R. C. Beavis. TANDEM: Matching Proteins with Tandem Mass Spectra. *Bioinformatics*, 20(9):1466–1467, 2004.

[9] D. L. Tabb, C. G. Fernando and M. C. Chambers. MyriMatch: Highly Accurate Tandem Mass Spectral Peptide Identification by Multivariate Hypergeometric Analysis. *Journal of Proteome Research*, 6(2):654–661, 2007.

[10] P. A. Pevzner, Z. Mulyukov, V. Dancik and C. L. Tang. Efficiency of Database Search for Identification of Mutated and Modified Proteins via Mass Spectrometry. *Genome Research*, 11(2):290–299, 2001.

[11] UNIMOD. Available at: `http://www.unimod.org/`, [cited June 24, 2013].

[12] S. R. Ramakrishnan, R. Mao, A. A. Nakorchevskiy, J. T. Prince, W. S. Willard, W. Xu, E. M. Marcotte and D. P. Miranker. A Fast Coarse Filtering Method for Peptide Identification by Mass Spectrometry. *Bioinformatics*, 22(12):1524–1531, 2006.

[13] D. Dutta and T. Chen. Speeding up Tandem Mass Spectrometry Database Search: Metric Embeddings and Fast Near Neighbor Search. *Bioinformatics*, 23(5):612–618, 2007.

[14] J. Novák, D. Hoksza, J. Lokoč and T. Skopal. On Optimizing the Non-metric Similarity Search in Tandem Mass Spectra by Clustering. In L. Bleris, I. Mandoiu, R. Schwartz and J. Wang (editors), *Proc. 8th International Symposium Bioinformatics Research and Applications*, volume 7292 of *Lecture Notes in Bioinformatics (LNBI)*, pages 189–200. 2012.

[15] J. Novák, T. Skopal, D. Hoksza and J. Lokoč. Non-metric Similarity Search of Tandem Mass Spectra Including Posttranslational Modifications. *Journal of Discrete Algorithms*, 13:19–31, 2012.

[16] X. Liu, A. Mammana and V. Bafna. Speeding up Tandem Mass Spectral Identification using Indexes. *Bioinformatics*, 28(13):1692–1697, 2012.

[17] R. Mao, S. R. Ramakrishnan, G. Nuckolls and D. P. Miranker. An Inverted Index for Mass Spectra Similarity Query and Comparison with a Metric-space Method: Case Study. In *Proceedings of the 3rd International Conference on Similarity Search and Applications (SISAP)*, pages 93–99. 2010.

[18] B. Lu and T. Chen. A Suffix Tree Approach to the Interpretation of Tandem Mass Spectra: Applications to Peptides of Non-specific Digestion and Post-translational Modifications. *Bioinformatics*, 19(suppl_2):ii113–ii121, 2003.

[19] C. Zhou, H. Chi, L.-H. Wang, Y. Li, Y.-J. Wu, Y. Fu, R.-X. Sun and S.-M. He. Speeding up Tandem Mass Spectrometry-based Database Searching by Longest Common Prefix. *BMC Bioinformatics*, 11(1):577, 2010.

[20] P. J. Ulintz, J. Zhu, Z. S. Qin and P. C. Andrews. Improved Classification of Mass Spectrometry Database Search Results Using Newer Machine Learning Approaches. *Molecular and Cellular Proteomics*, 5(3):497–509, 2006.

[21] D. C. Anderson, W. Li, D. G. Payan and W. S. Noble. A New Algorithm for the Evaluation of Shotgun Peptide Sequencing in Proteomics: Support Vector Machine Classification of Peptide MS/MS Spectra and SEQUEST Scores. *Journal of Proteome Research*, 2(2):137–146, 2003.

[22] K. Ning, H. K. Ng and H. W. Leong. PepSOM: An Algorithm for Peptide Identification by Tandem Mass Spectrometry based on SOM. *Genome Informatics*, 17:194–205, 2006.

[23] L. Wang, W. Wang, H. Chi et al. An Efficient Parallelization of Phosphorylated Peptide and Protein Identification. *Rapid Communications in Mass Spectrometry*, 24(12):1791–1798, 2010.

[24] Y. Li and X. Chu. Speeding up Scoring Module of Mass Spectrometry Based Protein Identification by GPU. In *2012 IEEE 14th International Conference on High Performance Computing and Communication & 2012 IEEE 9th International Conference on Embedded Software and Systems (HPCC-ICESS)*, pages 1315–1320. 2012.

[25] I. Bogdan, D. Coca, J. Rivers and R. J. Beynon. Hardware Acceleration of Processing of Mass Spectrometric Data for Proteomics. *Bioinformatics*, 23(6):724–731, 2007.

[26] C. Y. Park, A. A. Klammer, L. Käll, M. J. MacCoss and W. S. Noble. Rapid and Accurate Peptide Identification from Tandem Mass Spectra. *Journal of Proteome Research*, 7(7):3022–3027, 2008.

[27] B. J. Diament and W. S. Noble. Faster SEQUEST Searching for Peptide Identification from Tandem Mass Spectra. *Journal of Proteome Research*, 10(9):3871–3879, 2011.

[28] C. D. Wenger and J. J. Coon. A Proteomics Search Algorithm Specifically Designed for High-Resolution Tandem Mass Spectra. *Journal of Proteome Research*, 12(3):1377–1386, 2013.

[29] E. J. Hsieh, M. R. Hoopmann, B. MacLean and M. J. MacCoss. Comparison of Database Search Strategies for High Precursor Mass Accuracy MS/MS Data. *Journal of Proteome Research*, 9(2):1138–1143, 2010.

[30] L. Käll, J. D. Storey, M. J. MacCoss and W. S. Noble. Assigning Significance to Peptides Identified by Tandem Mass Spectrometry Using Decoy Databases. *Journal of Proteome Research*, 7(1):29–34, 2008.

[31] J. E. Elias and S. P. Gygi. Target-decoy Search Strategy for Increased Confidence in Large-scale Protein Identifications by Mass Spectrometry. *Nature Methods*, 4(3):207–214, 2007.

[32] J. Novák, T. Sachsenberg, D. Hoksza, T. Skopal and O. Kohlbacher. A Statistical Comparison of SimTandem with State-of-the-Art Peptide Identification Tools. In M. Mohamad, L. Nanni, M. Rocha and F. Fdez-Riverola (editors), *7th International Conference on Practical Applications of Computational Biology and Bioinformatics*, volume 222 of *Advances in Intelligent and Soft-Computing (AISC)*, pages 101–109. 2013. Available at: `http://www.simtandem.org` or `http://www.siret.cz/simtandem`, [cited June 24, 2013].

[33] M. Sturm, A. Bertsch, C. Gropl et al. OpenMS – An Open-source Software Framework for Mass Spectrometry. *BMC Bioinformatics*, 9(1):163, 2008.

[34] O. Kohlbacher, K. Reinert, C. Gröpl, E. Lange, N. Pfeifer, O. Schulz-Trieglaff and M. Sturm. TOPP – the OpenMS Proteomics Pipeline. *Bioinformatics*, 23(2):e191–e197, 2007.

[35] M. Beck, A. Schmidt, J. Malmstroem, M. Claassen, A. Ori, A. Szymborska, F. Herzog, O. Rinner, J. Ellenberg and R. Aebersold. The Quantitative Proteome of a Human Cell Line. *Molecular Systems Biology*, 7:549, 2011.

[36] UniProtKB/Swiss-Prot. Available at: `http://www.uniprot.org/`, [cited June 24, 2013].

[37] MSDB. Available at: `http://ftp.ncbi.nih.gov/repository/MSDB/`, [cited June 24, 2013].

[38] NCBI RefSeq. Available at: `http://www.ncbi.nlm.nih.gov/RefSeq/`, [cited June 24, 2013].

[39] S. Nahnsen, T. Sachsenberg and O. Kohlbacher. PTMeta: Increasing Identification Rates of Modified Peptides using Modification Prescanning and Meta-analysis. *Proteomics*, 13(6):1042–1051, 2013.

[40] Y. Fu. Bayesian False Discovery Rates for Post-translational Modification Proteomics. *Statistics and Its Interface*, 5:47–59, 2012.