# An Application of the Metric Access Methods to the Mass Spectrometry Data

Jiří Novák and David Hoksza

***Abstract*— Mass spectrometry is a very popular method for protein and peptide identification nowadays. Abundance of data generated in this way grows exponentially every year and although there exist algorithms for interpreting mass spectra, demand for faster and more accurate approaches remains.**

**We propose an approach for preprocessing the protein sequence database based on metric access methods. This approach allows to select only a small set of suitable peptide sequence candidates, which can be then compared with experimental spectra using more sophisticated algorithms. We define logarithmic distance for selecting peptide sequence candidates and also outline possibilities of using the interval query for searching posttranslational modifications.**

**The experimental results show that our approach is comparable in precision with nowadays most widely used public tools and outline possible directions for further resarch.**

## I. Introduction

MASS SPECTROMETRY [10] is a modern and fast method for protein sequences identification. The knowledge of protein sequences is important for studying protein functions, that are based on protein structure determinded by the sequences [20]. Mass spectrometry is also denoted as an indirect sequencing, because the mass spectrometer does not determine a linear sequence of amino acids (protein's primary structure), but it captures a collection of uninterpreted data called mass spectrum. Subsequently, the protein sequence must be elucidated from the mass spectrum[1] by some sophisticated algorithm first.

Before the mass spectrometry analysis, an unknown protein sample (number of molecules with identical structure) is digested by a specific enzyme into many shorter peptides. After the protein digestion at specific cleavage sites the sample is analyzed by the mass spectrometer. Single peptide molecules get charges (becoming to be ions) and they are separated by their ratios mass/charge ($m/z$).

In case of simple mass spectrometry (MS), the mass spectrum is a list of detected ratios $m/z$ with intensities of their occurence (list of peaks). Moreover, in case of tandem mass spectrometry (MS/MS), each peptide ion can be splitted to various types of fragment ions, hence for one analyzed protein we get a collection of the tandem mass spectra (one spectrum for each peptide ion) [15]. The peptide ion common to all fragment ions in a tandem spectrum is denoted as parent peptide ion. The most common types of fragment ions $y$, $b$ and $a$ are presented in the Fig. 1.

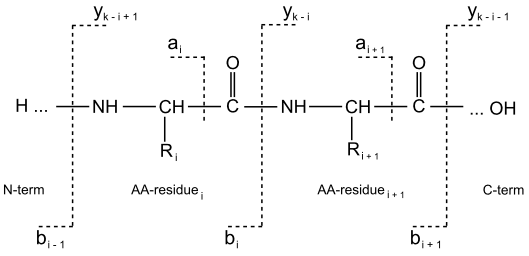


Fig. 1. The most common types of peptide fragment ions in a tandem mass spectrum (AA = amino acid).

Two basic approaches are used for interpreting mass spectra. First is based on the direct spectra interpretation using graph algorithms (De Novo peptide sequencing) [7] and it is typical for tandem mass spectra. Second is based on scanning databases of already known protein or peptide sequences, which is denoted Peptide Mass Fingerprinting (PMF) [9] for MS spectra and Peptide Fragment Fingerprinting (PFF) [13] for MS/MS spectra. A combined approach Sequence Tag [16] can be used for tandem mass spectra, where a short aminoacid sequence (tag) is determined manually or using graph algorithm first and then the database is searched.

The interpretation of experimental spectra is complicated, because many peaks (up to 80%) cannot be usually recognized. These unrecognizable peaks (called noise) correspond to ions with unpredictable structure. The noise arises from ions losing complex chemical groups, from ions falling into many parts or it can be a consequence of admixtures in the analyzed sample, etc. Moreover, not all of the theoretically generated ions (e.g. y or b-ions which are the most important ones for peptide identification) need to have their counterparts in the tandem mass spectra. Previous obstacles (noise, absence of some important peaks) makes De Novo identification very difficult, if not in practice impossible, task. On the other hand, a substantial drawback of the database approach is the need of existence of the analyzed or related sequence in the database.

Databases often result from translation of known DNA sequences, hence unknown protein sequences can be identified. The identification of protein sequences is complicated, if the proteins were modified after translation (posttranslation modifications, PTM) or during preparation the sample for mass analysis, because the peaks can be shifted.

Jiří Novák, Czech Technical University in Prague, Faculty of Electrical Engineering, Czech Republic, email: jirinovak@atlas.cz and David Hoksza, Charles University in Prague, Faculty of Mathematics and Physics, Czech Republic, email: david.hoksza@mff.cuni.cz.

This work was supported by the Grant Agency of Charles University under Project Code GAUK 57907 and GACR grant under project code 201/09/0683.

[1]In the case of tandem mass spectrometry from the collection of the mass spectra.

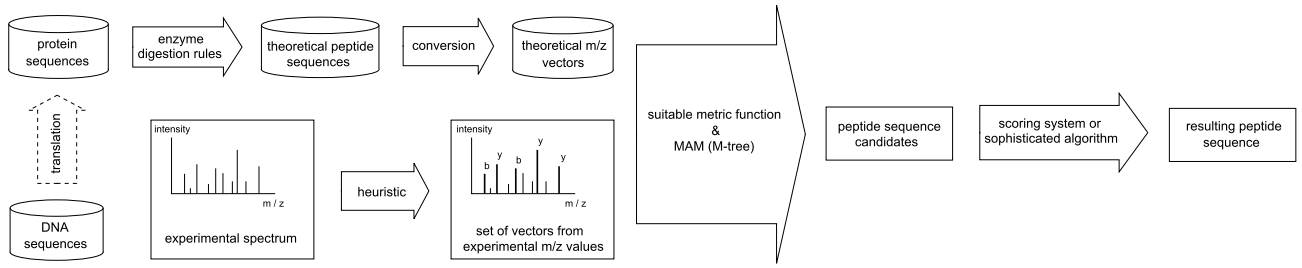


Fig. 2. Basic principles of proposed method.

The abundance of identified sequences in bioinformatic databases is growing exponentially every year [11], so the database should be preprocessed to speed up the search. An approach based on the metric access methods (MAMs) [23] is proposed in this work. Although the M-tree [6] structure is used in the experimental part, any other MAM can be utilized. The entire identification algorithm can be labeled as PFF, because it focuses on the interpretation of the tandem mass spectra and on the searching in protein sequences databases.

Basic principles of proposed method are presented in Fig. 2. The vectors of theoretical $m/z$ values are generated from already known peptide sequences and the vectors of experimental $m/z$ values are selected from an experimental spectrum by heuristic. For comparison of theoretical and experimental vectors suitable metric functions are used together with M-tree for speeding up the search. On a small set of selected peptide sequence candidates, a simple scoring system or some more sophisticated algorithm can be applied to select a sequence best corresponding to the experimental spectrum. Possibilities for searching peptide modifications are also outlined (section II-E).

Our proposed method is based on using lowdimensional metric spaces, but a method based on highdimensional metric spaces was also presented in [8]. High dimension can cause problems when using MAMs like M-tree, since the range and k-NN ($k$-nearest neighbors) query algorithms have an exponential dependency on the dimension of the metric space. Another existing method [22] is based on reducing the intrinsic dimensionality of the highdimensional spaces and using MVP-tree [23].

The rest of this paper is organized as follows: Section II describes the method for generating database of vectors from protein sequences, heuristics for selecting vectors from experimental spectra and metrics for their comparing. There is also presented a simple scoring system and an algorithm for searching peptide modifications. Section III introduces the M-tree and the interval query. Experiments and comparison with public available searching tools MASCOT MS/MS Ions Search [1] and ProteinProspector MS-Tag [2] are presented in Section IV to show the effectivity of used metrics and the M-tree performance. Finally, Section V concludes the paper.

## II. Methods

### A. Database construction

Before the mass analysis each protein sequence is digested at the specific positions (cleavage sites) to many shorter peptides by an enzyme e.g. trypsine[2] [18]. So, the theoretical peptide sequences and their theoretical fragment ions spectra[3] can be generated for each protein sequence. The minimum and maximum size of generated peptide sequences and the maximum number of missed cleaveage sites[4] must be also defined with enzyme type.

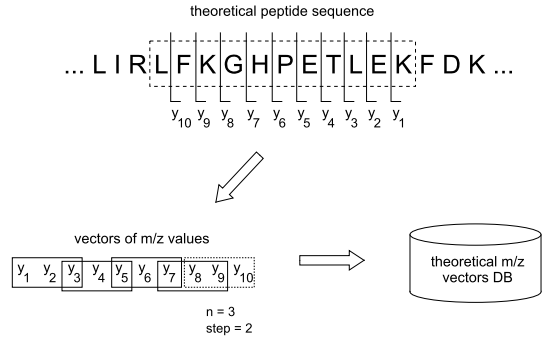


Fig. 3. Database construction.

Many types of fragment ions (e.g. $y$, $b$, $a$, $y$-$H_2O$, $y$-$NH_3$, $y$-$H_2O$-$NH_3$, $b$-$H_2O$, $a$-$H_2O$) can be generated into the theoretical fragment ion spectrum. But bigger set of fragment ions means higher storage requirements in database without certainty that all theoretically generated ions will also be present in the experimental spectrum. Thus, we will generate and store for each peptide only its $y$-ions, eventually $y$ and $b$-ions (because these are the most abundant ones).

Since the theoretical spectra have variable number of peaks and we need vectors of constant size for a MAM method, a sliding window of size $n$ is applied. The window is moved over the theoretical spectrum[5] with specified step and vectors of $m/z$ values bounded by this window are stored in the

[2]Trypsine digests protein sequence after aminoacid residues lysine (K) and arginine (R), if they are not followed by proline (P).

[3]The fragment ions intensities cannot be known by generating the theoretical spectrum, so only $m/z$ values are used.

[4]The digestion by enzyme is not perfect in reality, thus some digestion places can be missed.

[5]Over the $m/z$ values sorted in ascending order.

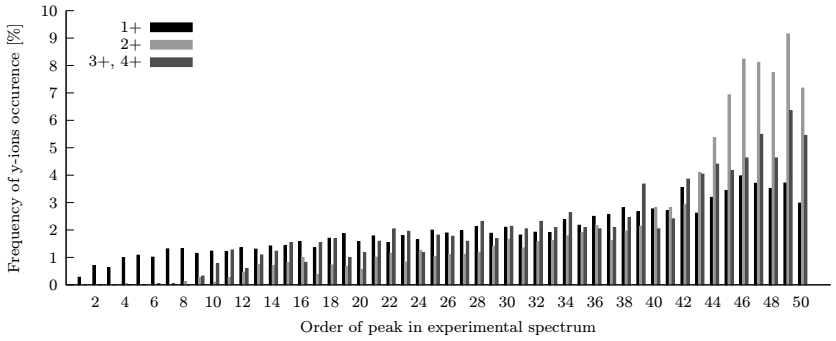


Fig. 4. Heuristic based on fragment ions positions (collection *amet* - see section IV; experimental spectra are differentiated by parent peptide charge and aligned against themselves from the right).

database (Fig. 3). If $step < n$, the redundantion of indexed data increases, but also precision of peptide identification improves. If there exists a peak not being member of a window, a window for last $n$ values is added.

### B. Heuristics

The intensity of fragment ions cannot be known for theoretical spectra and the experimental spectra does not guarantee, that peaks with lower intensity are less interesting for peptide identification than peaks with higher intensity. So, we propose two heuristics for selecting the set of peaks from an experimental spectrum without dependency on intensity values. These peaks ($m/z$ values) form the vectors needed for search the database.

The first heuristic is based on a simply idea of $y$ fragment ions positions. If the database is constructed only from $y$-ions, the vector created from an experimental spectrum should consist in the ideal case of $y$-ions only (another included ions can be understood as noise). The probability of selecting a continuous vector consisted only of $y$-ions is higher at the end[6] of the experimental spectra (especially in spectra with parent peptide charge $z \geq 2^+$), because the $y$-ions are often cumulated there (Fig. 4). So, the heuristic based on the idea of $y$ fragment ions positions picks the last $n$ peaks from the end of spectrum for indexing.

With increasing vector dimension, probability of vector continuity decreases. Since the experimental spectrum vector does not need to be found in the theoretical peptide spectrum, we set a sliding window on the last $n$ peaks of experimental spectra and repeatedly move it towards the begin of the spectrum with the specified step.

Let the fragment ions $b_i$ and $y_{k-i}$ be complementary for peptide sequence of size $k$, where $0 < i < k$ (Fig. 1). The second heuristic is based on the idea of searching complementary $b$ and $y$-ions. The sum of masses $b_i$ and complementary $y_{k-i}$ ion is equal[7] to the mass of neutral peptide molecule $m_p$ plus 2 Daltons (1).

$$m(b_i) + m(y_{k-i}) = m_p + 2 \quad (1)$$

Because it cannot be decided which peak correspond to $b$-ion and which to $y$-ion, a following procedure can be used. When we suspect a mass $m$ of being $b$-ion then the spectrum could also contain masses coresponding $a$-ion, $a$-ion after losing water $H_2O$ or ammoniac $NH_3$. If so, $m$ is considered $b$-ion and complementary mass to $y$-ion.

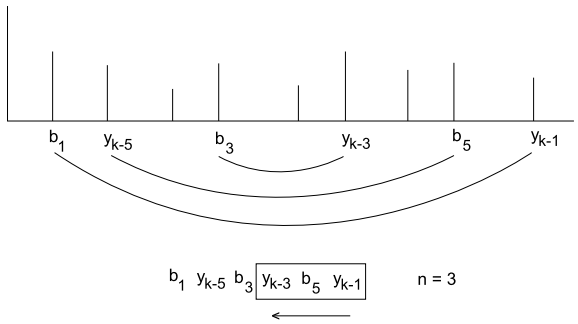


Fig. 5. Heuristic based on searching complementary $b$ and $y$-ions.

The process is similar to the first heuristic, except for the database vectors contain $b$ and $y$-ions this time and the last $n$ peaks marked by this heuristic are selected from the experimental spectrum. The sliding window can be moved over the marked $b$ and $y$-ions to increase effectivity of the search (Fig. 5).

The heuristic based on searching complementary $b$ and $y$-ions is most effective for spectra from collection *amet* with parent peptide charge $1^+$. Probability that a peak marked by this heuristic really correspond to $b$ or $y$-ion is about $90\%$, but only $40\%$ of spectra have at least 10 marked peaks by this heuristic (Fig. 6). More marked peaks means possible higher maximum number of shifts of the sliding window, hence higher probability of match.

The heuristic can recognize about $55\%$ of all peaks corresponding to $b$ or $y$-ions in an experimental spectrum without differentiation them and about $22\%$ peaks with differentiation [17]. A drawback of differentiation is its low precision, hence further we use an alternative where we do not differentiate $b$ and $y$-ions.

[6]The implicit ordering of the peaks by increasing $m/z$ values is assumed in an experimental spectrum.

[7]Providing that the charge $z = 1^+$ and specified mass tolerance is reflected.

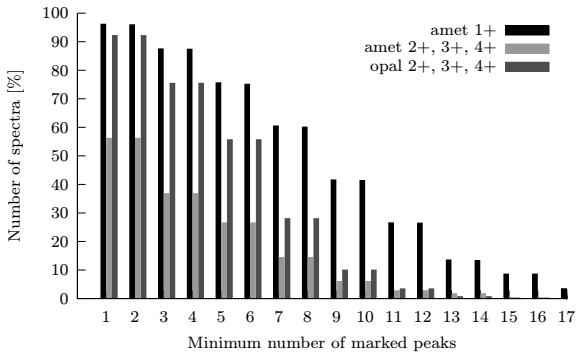


Fig. 6. Minimum of selected peaks using heuristic based on searching complementary $b$ and $y$-ions.

### C. Metrics

Metrics suitable for comparing theoretical and experimental vectors are described below. A metric $d$ defines the similarity of two objects and satisfies following properties: reflexivity, positiveness, symmetry and triangle inequality. If the triangle inequality is not satisfied, $d$ is called semimetric. When the positiveness property is not satisfied, $d$ is called pseudometric and it can be turned into metric by treating each pair of objects as a single object (value of $d$ is 0) [23].

Well known metric, which falls into group of Minkowski distances, is maximum distance (2).

$$L_\infty(\vec{x},\vec{y}) = \max_{i=1}^{n} ||x_i - y_i|| \tag{2}$$

Another, for our purpose interesting metric, is the Hausdorff distance (3). It first computes distances to nearest neighbours in set $B$ for all elements from set $A$ and for all the elements from set $B$ distances to their respective nearest neighbours in set $A$. Finally, from computed distances the maximum value is selected. The inner function $d_x$ must be a metric. Further we treat $d_x$ as difference of two points in Euclidean space.

$$\begin{aligned} h(A,B) &= \max_{a\in A}\left\{\min_{b\in B}\left\{d_x(a,b)\right\}\right\} \\ h(B,A) &= \max_{b\in B}\left\{\min_{a\in A}\left\{d_x(a,b)\right\}\right\} \\ H(A,B) &= \max(h(A,B),h(B,A)) \end{aligned} \tag{3}$$

Because the elements with minimum distances are chosen, there is a particular resistance against shifts of peaks ($m/z$ values) with small differences. For example, vectors $\vec{x} = \{300, a, 400, 500\}$ and $\vec{y} = \{300, 400, 500, b\}$ have always $L_\infty \geq 100$, for $300 < a < 400$ and $b > 500$. On the other hand for $a = 399$ and $b = 501$, $H(\vec{x},\vec{y}) = 1$.

The cosine similarity (4) is also presented in mass spectrometry literature [14]. Cosine of an angle is not a metric, but utilize the $\arccos$ function and operate directly with the angle size, which is metric.

$$\cos\theta = \frac{\vec{x}\vec{y}}{\|\vec{x}\|\,\|\vec{y}\|} \tag{4}$$

Majority of metrics have shortcoming presented in following example. Lets assume vectors $\vec{x} = \{200, 300, 400, 500\}$, $\vec{y} = \{200, 300, 460, 500\}$ and $\vec{z} = \{210, 305, 420, 475\}$. The vectors $\vec{x}$ and $\vec{y}$ are closer considering peptide identification, but their Euclidean distance of the vectors $\vec{x}$ and $\vec{z}$ is lower. Such situations are the reason for proposing a modified distance which we call *logarithmic distance* (5).

$$\begin{aligned} d_l(x_i,y_i) &= \log|x_i - y_i|, & |x_i - y_i| > 1 \\ &= 0, & otherwise. \end{aligned}$$

$$d_{log}(\vec{x},\vec{y}) = \sum_{i=1}^{n} d_l(x_i,y_i) \tag{5}$$

The logarithm function causes that $\vec{x}$ and $\vec{y}$ are closer than $\vec{x}$ and $\vec{z}$, but it disadvantages also the vectors with a small error constant. For example, vectors $\vec{x'} = \{200, 300, 400, 500\}$ and $\vec{y'} = \{210, 310, 410, 510\}$ can represent the same or similar peptide sequences, because the small constant errors can be consequence of an aminoacid mass modification, an aminoacid mutation or the whole spectrum contamination (e.g. by sodium ions $Na^+$).

The logarithmic distance is not a metric, since it does not satisfy the positiveness and triangle inequality, so it is only a pseudo-semimetric. But for the mass spectrometry data (processed as described in sections II-A and II-B) is triangle inequality satisfied with high probability (nearly 100%) and so this distance is good in practice [17].

In case of troubles with inequality is also possible work with modified triangle inequality (6), where constant $\kappa \geq 0$ can be determined empirical. But using of this alternative get worse efficiency by using the metric access methods.

$$d(a,b) + d(b,c) + \kappa \geq d(a,c) \tag{6}$$

### D. Searching and scoring

If the database of vectors was created (section II-A) and some metric was defined (section II-C), then the M-tree can be constructed. If M-tree was created, the set of vectors corresponding to peptide sequence candidates is selected easily using the range query (section III) and a query vector obtained by a heuristic (section II-B). If the correct peptide sequence was not found among the candidates another query vector is used and the search is repeated. The query vectors are specified by a sliding window (section II-B) and maximum number of window shifts is defined by the user. Exceeding maximum number of shifts means, the peptide sequence was not found for a given experimental spectrum.

For the set of peptide sequence candidates obtained by the range query, a simply scoring system is employed. For each candidate sequence the theoretical $b$ and $y$-ions are generated. If the sequence candidate with maximum $b$ and $y$-ions matched in experimental spectrum corresponds to the reference sequence listed in experimental data collection, the correct peptide sequence is matched. This scoring system can be easily replaced with a more sophisticated algorithm, such as spectral alignment based on dynamic programming [12].

### E. Searching modifications

Searching peptide modifications is still an open problem [21]. A simple method is based on adding modified peptide sequences to the database, with the drawback that the number of possible modifications is immense [4]. An algorithm based on using a set of interval queries is proposed below.

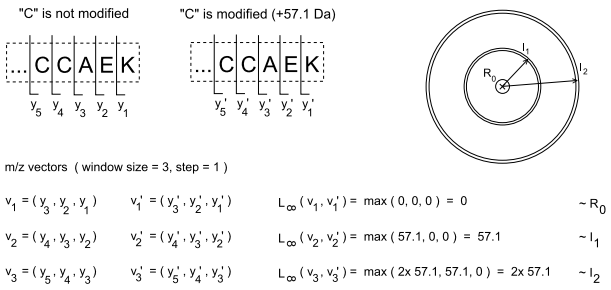


Fig. 7. Using a set of interval queries for searching peptide modifications.

In addition to one range query for searching unmodified peptide sequence, a set of interval queries for searching peptide modifications is used (Fig. 7). Suitable metrics for this purpose are those that do not cummulate errors e.g. maximum (2) or Hausdorff distance (3), because each interval query corresponds to an error caused by an aminoacid mass modification in Daltons. The search for one modification can correspond to more interval queries in case of its repeated occurence in the peptide. If the sequence with at most three occurences of carbamidated cysteine (+57.01 Da) was searched, the three interval queries $I(q, r-tol, r+tol)$, would have to be used, where $r = \{57.01, 2 \times 57.01, 3 \times 57.01\}$ and $tol$ is the mass tolerance.

A problem can arise with increasing number of searched peptide modifications, because of the exponential growth of required interval queries. If we grant $m$ repeated occurences of $n$ different aminoacid modifications, then the number of interval queries corresponds to the sum of combinations with repeated elements (7). For searching e.g. $n = 3$ different modifications, we use 3 interval queries for $m = 1$ occurences, 9 for $m = 2$ and 19 for $m = 3$. The number of known aminoacid modifications can be counted for hundreds at this time [4]. Fortunately, only few modifications, that usually appear in experimental spectra, can be predicted empirical.

$$\sum_{k=1}^{m} \left( \begin{array}{c} n+k-1 \\ k \end{array} \right) \tag{7}$$

The scoring algorithm (section II-D) must be also modified by using the set of interval queries. The theoretical values of $b$ and $y$-ions with shifts of masses caused by searched modifications are generated. For example, if we want to search one occurence of oxidized methionine (+16 Da) together with one occurence of carbamidated cysteine (+57.01 Da), then the interval query average radius is $r_{avg} = 16 + 57.01$. The additions +16 Da for methione (M) and +57.01 Da for cysteine (C) are applied to the theoretical $b$ and $y$-ions. The sequence best corresponding to the experimental spectrum stays the sequence with highest number of matched $b$ and $y$-ions.

Alg. 1. Searching modifications

```
1  best = ∅;
2  while not exceeded max. # of window shifts {
3    query = vector obtained by heuristic;
4    result = rangeQuery(query,radius);
5    best_one = scoring(result);
6    if best_one.better(best) {
7      best = best_one;
8      if best.isGuaranteed() return best; }
9    for all analyzed modifications {
10     result = intervalQuery(query,
11              modif.mass()-tol,
12              modif.mass()+tol);
13     best_one = scoring(result,modif);
14     if best_one.better(best) {
15       best = best_one;
16       if best.isGuaranteed() return best; } } }
17 return not found;
```

## III. Metric access methods

The metric access methods (MAMs) [23] have been designed to quickly search in databases modeled in metric spaces. They use the triangle inequality to organize data objects into metric regions and for pruning those regions. MAM of our choice for experimental evaluation is the metric tree (M-tree) [19].

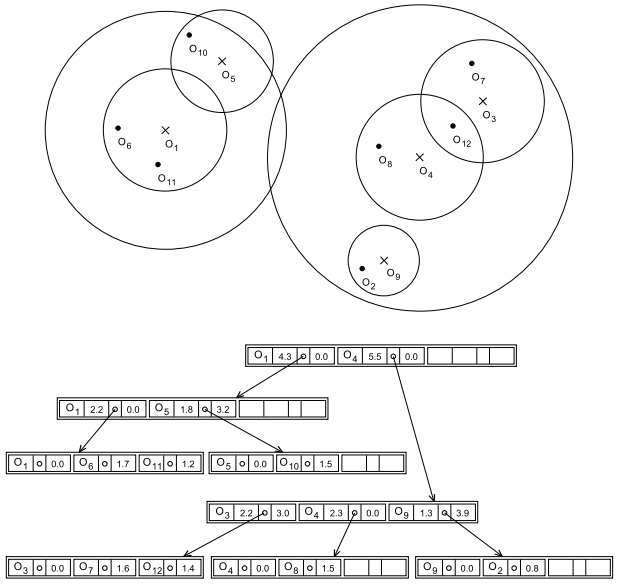

Fig. 8. The M-tree structure.

The M-tree is a dynamic, hierarchical and balanced index structure. It contains $n$-dimensional hyper-spherical regions storing $n$-dimensional vectors representing the indexed objects. Some objects are selected as pivots (centers of hyperspherical regions), the others are partitioned among these regions (Fig. 8).

Several types of queries can be performed by MAMs. First and foremost the range (8) and the $k$-nearest neighbors (k-NN) [23] queries.

$$R(q, r) = \{o \in X, d(o, q) \leq r\} \tag{8}$$

The interval (or hyper-ring) query (9) is defined as an extension of the range query for searching peptide modifications. The average radius $r_{avg}$ is the average value of $r_{min}$ and $r_{max}$. The M-tree interval query algorithm is proposed in [17].

$$I(q, r_{min}, r_{max}) = I(q, r_{avg} - t, r_{avg} + t) = \\ = \{o \in X, d(o, q) \geq r_{min} \wedge d(o, q) \leq r_{max}\} \tag{9}$$

## IV. Experiments

We use Amethyst and Opal collections of experimental spectra in our experiments, which are part of Quartz project presented on the GPM site (The Global Proteome Machine Organization) [3]. The collections are formed from tandem mass spectra of peptides founded in human genome. Datasets *amethyst-gv.xml* (*amet*) and *opal-gv.xml* (*opal*) are used in the experiments (Tab. I). We assume, that spectra with parent peptide charge $1^+$ are the tandem MALDI (Matrix Assisted Laser Desorption Ionization) [5] spectra and the others are the ESI (Electrospray Ionization) [10] spectra. We focus on interpretation of single peptide sequences from tandem mass spectra.

| Collection | Count of spectra | | | | |
|---|---|---|---|---|---|
| | Total | $1^+$ | $2^+$ | $3^+$ | $4^+$ |
| *amet* | 1825 | 1052 | 533 | 224 | 16 |
| *opal* | 622 | 0 | 477 | 139 | 6 |

TABLE I
THE DATA COLLECTIONS.

The protein sequence database used in experimental part was created from all protein sequences referenced in *amet* and *opal* collections. In order to increase the database size, we enriched it by adding extra human protein sequences and removed duplicate sequences. Parameters for generating database from protein sequences were - enzyme: trypsine; minimum peptide size: 6; maximum peptide size: 20; maximum of missed cleavage sites: 1; sliding window step: 1 [8].

All experiments were performed on machine with dualcore 64-bit processor AMD TURION TL52, 120 GB HDD, 1 GB RAM and OS Windows XP SP2. The time of computation, number of readed nodes and selectivity are average values per one spectrum implicitly.

### A. Distances comparison

The Euclidean, maximum, Hausdorff and logarithmic distance and cosine similarity were compared. The database was constructed from $y$-ions and the heuristic based on $y$-ions positions was used (section II). Only the last $n$ peaks from each experimental spectrum were selected, where $n$ is a vector dimension.

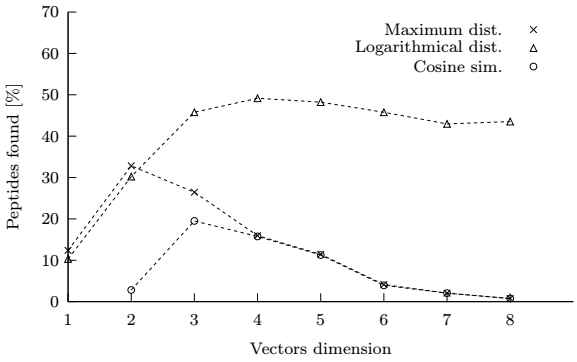


Fig. 9. Distances comparison (collection: *amet*; parent peptide charge: $2^+$; query type: 20-NN; peptides in database: $\approx$ 20000).

The most peptide sequences were identified with the logarithmical distance (Fig. 9). The quality of identification slowly decreases with increasing dimension for this distance. Euclidean, maximum and Hausdorff distances give almost identical results (only the maximum distance is shown in Fig. 9). The quality of identification decreases quickly with increasing dimension in their case and for $n > 5$ is too small for practical use, but they still give a little better results than cosine similarity.

### B. Searching peptide modifications

The search for peptide modifications was tested using a set of interval queries and the maximum distance. One range query with $r = 10$ and 8 interval queries with average radiuses comming from the set $\{57.01, 2 \times 57.01, 16, 17.016, 14.0156, 1, 30.0106, 57.01 + 42\}$ were used. The radiuses correspond to modifications of aminoacids {C, 2×C, M, Q, V $\vee$ S $\vee$ A $\vee$ T, N $\vee$ Q, G, C + D}, where the mass tolerance needed for specifying the radiuses $r_{min}$ and $r_{max}$ is $\pm 0.1$ Da. The database was constructed from $y$-ions and heuristic based on $y$-ions positions was used. Fig. 10 shows the dependency of recognized proteins on increasing number of window shifts.

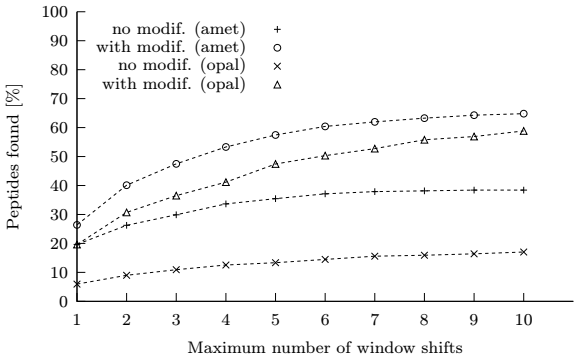


Fig. 10. Searching peptide modifications (vectors dimension: 3; parent peptide charge: $\geq 2^+$; peptides in database: 23326).

[8]With bigger step, worse results were obtained.

The quality of recognition increases with growing maximum number of window shifts. The increase is most evident for number of window shifts $\leq 5$. For *amet*, the number of identified peptides is $1.69\times$ higher when searching modifications were included than without it. The quality of identification reaches to $64.81\%$, for collection *opal* is count $3.45\times$ higher and the quality reaches $58.84\%$ (10 window shifts).

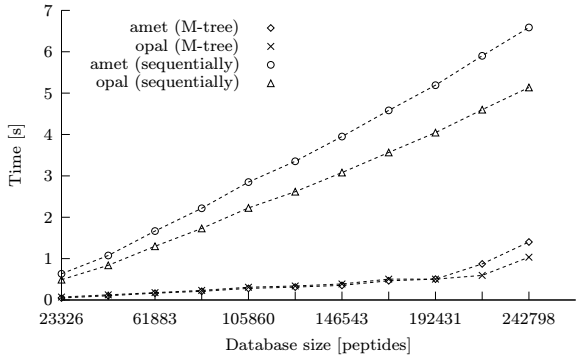


Fig. 11. Time of identification (vectors dimension: 3; parent peptide charge: $\geq 2^+$; maximum number of window shifts: 10; type of searching: with modifications; inner nodes capacity: 50 items; leaf nodes capacity: 66 items).

Databases containing from 510 to 5468 human protein sequences (from 23326 to 242798 peptide sequences) were used for testing M-tree structure's characteristics. The time of identification is much lower when using M-tree than with sequential algorithm (Fig. 11). Average speed up is $9.82\times$ for *amet* collection and $7.19\times$ for *opal* collection. The relative number of fetched M-tree nodes quickly decreases with increasing size of the database. It decreased for both collections by using $10.5\times$ bigger database from around $65\%$ to $34\%$. The results on the real data (consisting "all" human proteins) are presented in section IV-D.

### C. Logarithmic distance

Logarithmic distance was tested using heuristic based on searching complementary $b$ and $y$-ions (section II-B). Results obtained by the second heuristic are presented in section IV-D. Since we do not have an effective technique for distinguishing $b$ and $y$-ions (section II-B), the ions were stored in the database without differentiating them. Since logarithmic distance cummulates errors, no searching modifications were applied. Dimension of vectors was $n = 5$, hence only those spectra where the heuristic marked at least 5 peaks were selected for further identification.

The quality of identification and the selectivity in dependency on radius of M-tree range query are presented in Fig. 12. The range query radius increases from $r = 1.2$ to $r = 4.8$. The quality of identification grows from $18.62\%$ to $44.28\%$ and the selectivity from $0.00018\%$ to $0.23366\%$. Finally, the number of readed nodes increases from $9.52\%$ to $43.71\%$.

The datasets containing from 510 to 4476 human protein sequences (from 23326 to 192431 peptide sequences) were also used for testing characteristics of logarithmical distance. The identification was average $11.27\times$ faster by using M-tree than by using sequential algorithm (Fig. 13). The number of readed nodes decreased from $29.06\%$ to $18.50\%$. The average selectivity was about $0.01\%$.

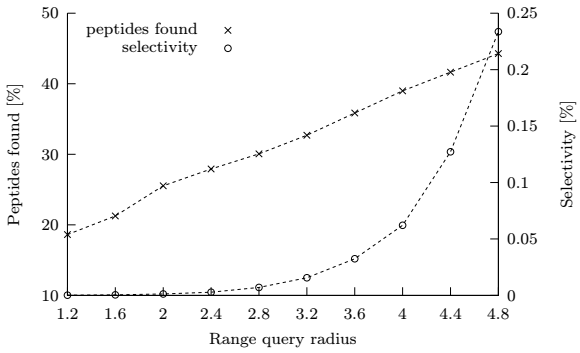


Fig. 12. Range query (collection: *amet*; parent peptide charge: $1^+$; number of spectra: 795; vectors dimension: 5; peptides in database: 61883; inner nodes capacity: 50 items; leaf nodes capacity: 63 items).

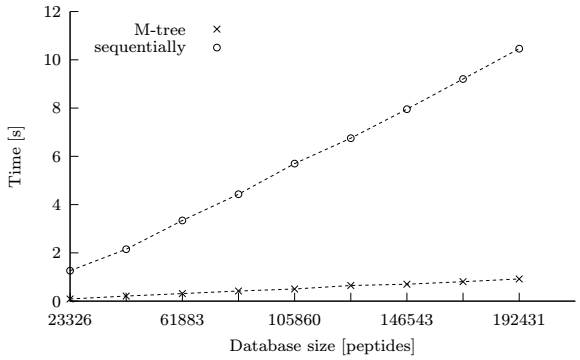


Fig. 13. Time of identification (collection: *amet*; parent peptide charge: $1^+$; number of spectra: 795; vectors dimension: 5; range query radius: 3.0; inner nodes capacity: 50 items; leaf nodes capacity: 63 items).

### D. Comparison with existing methods

Proposed methods were compared with free available searching tools MASCOT MS/MS Ions Search [1] and ProteinProspector MS-Tag (version 4.27.2 Basic) [2]. First, 50 MS/MS ESI-QTOF spectra from collection *opal* with parent peptide charge $2^+$ were analyzed. The identified peptide sequences were compared with referenced sequences in the data collection.

For MASCOT search tool following setup was used - database: SwissProt; taxonomy: human; enzyme: trypsine; missed cleavages: 1; fixed modifications: Carbamidomethyl C ($+57.01$ Da); variable modifications: Oxidation M ($+16$ Da), Gln$\rightarrow$pyro-Glu for N-term Q ($-17.016$ Da) and Deamidated N, Q ($+1$ Da); peptide charge: $2^+$; mass: monoisotopic; peptide mass tolerance: 1.2 Da; fragments mass tolerance: 0.2 Da. Settings for ProteinProspector were identical [17].

We employed maximum and Hausdorff distances in the algorithm for searching peptide modifications, while logarithmic distance was used in algorithm that did not used peptide modification. The database held 47781 human protein sequences (2427652 peptide sequences) in order to simulate free available searching tools. The heuristic based on $y$-ions positions was used.

Following modifications of aminoacids were searched for - $\{$C, M, Q, N $\vee$ Q$\}$ corresponding to average radiuses of the set of interval queries $\{57.01, 16, 17.016, 1\}$. Radiuses for interval queries were defined with tolerance $\pm 0.1$ for each average value. Other settings - vectors dimension: 3; M-tree inner nodes capacity: 50 items; M-tree leaf nodes capacity: 66 items; maximum of window shifts: 10; range query radius for maximum and Hausdorff distance: 10.0; range query radius for logarithmical distance: 2.0.

| **Searching type** | max. dist. + int. queries | Haus. dist + int. queries | log. dist. |
|---|---|---|---|
| **Peptides found** | 26 (52%) | 28 (56%) | 24 (48%) |
| **Search time (M-tree)** | 149.1 ms | 165.3 ms | 263.6 ms |
| **Search time (seq.)** | 295.5 s | 357.9 s | 157.6 s |
| **Time of indexing** | 31.1 ms | 34.3 ms | 23.6 ms |
| **Nodes read** | 3.32% | 3.80% | 11.15% |
| **Selectivity** | 0.27% | 0.28% | 0.10% |

TABLE II
COMPARING WITH EXISTING METHODS.

The MASCOT search tool identified 34 (68%) peptide sequences and guaranteed the results for 31 (62%) of them. The ProteinProspector identified 36 (72%) peptide sequences. The detailed statistics for M-tree are proposed in the table II (searching times, nodes read and selectivity are average values per one spectrum, time of indexing is average value per one peptide sequence). Times of identification could not be compared with MASCOT and ProteinProspector, because these informations were not available.

The number of identified peptides by M-tree is comparable with existing search tools, even thought simple heuristic for selecting vectors from experimental spectra and simple scoring system were used. The number of readed nodes is for alternative with maximum or Hausdorff distance a little lower than for logarithmical distance. The speed up of M-tree against the sequential algorithm is around $10^3$.

## V. CONCLUSIONS AND FUTURE WORK

An approch for preprocessing the proteins or peptides sequence database using metric access methods was proposed in this work. The logarithmical distance was defined and an algorithm for searching peptide modifications using set of interval queries was introduced. Finally, two heuristics were designed for creating the query objects from experimental spectra. First is based on $y$-ions positions, the other on searching complementary $b$ and $y$-ions.

The quality of identification is slightly worse in comparison to nowadays most widely used search engines but the overall setup of our algorithm enables using more sophisticated heuristics, new metrics and scoring shemes definitely open possibilities for a further research.

## REFERENCES


[1] MASCOT. http://www.matrixscience.com/.
[2] ProteinProspector. http://prospector.ucsf.edu/.
[3] The Global Proteome Machine Organization. http://www.thegpm.org/.
[4] Unimod. http://www.unimod.org/.
[5] P. Chaurand, F. Luetzenkirchen and B. Spengler, "Peptide and protein identification by matrix-assisted laser desorption ionization (MALDI) and MALDI-post-source decay time-of-flight mass spectrometry," *Journal of the American Society for Mass Spectrometry*, vol. 10, no. 2, pp. 91-103. 1999.
[6] P. Ciaccia, M. Patella and P. Zezula, "M-tree: An Efficient Access Method for Similarity Search in Metric Spaces," *Proc. of 23rd Int. Conf. on VLDB*, pp. 426-435. 1997.
[7] V. Dančík, T.A. Addona, K.R. Clauser, J.E. Vath and P.A. Pevzner, "De Novo Peptide Sequencing via Tandem Mass Spectrometry," *Journal of Computational Biology*, vol. 6, no. 3, pp. 327-342. 1999.
[8] D. Dutta and T. Chen, "Speeding up tandem mass spectrometry database search: metric embeddings and fast near neighbor search," *Bioinformatics Oxford Journal*, vol. 23, no. 5, pp. 612-618. 2007.
[9] W.J. Henzela, C.Watanabea and J.T. Stults, "Protein identification: the origins of peptide mass fingerprinting," *Journal of the American Society for Mass Spectrometry*, vol. 14, no. 9, pp. 931-942. 2003.
[10] E. Hoffmann and V. Stroobant, *Mass spectrometry: principles and applications.* 3rd ed., John Wiley & Sons, Chichester, England. 2007.
[11] D. Hoksza and T. Skopal, "Index-based approach to similarity search in protein and nucleotide databases," *CEUR Proc. Dateso 2007*, vol. 235, pp. 67-80. 2007.
[12] N.C. Jones and P.A. Pevzner, *An Introduction to Bioinformatics Algorithms.* MIT Press, Cambridge, Massachusetts. 2004.
[13] M. Kinter and N.E. Sherman, *Protein Sequencing and Identification Using Tandem Mass Spectrometry.* John Wiley & Sons, New York, USA. 2000.
[14] J. Liu, A.W. Bell, J.J.M. Bergeron, C.M. Yanofsky, B. Carrillo, C.E.H. Beaudrie and R.E. Kearney, "Methods for peptide identification by spectral comparison," *Proteome Science*, vol. 5, no. 3. 2007.
[15] R. Matthiesen, *Mass Spectrometry Data Analysis in Proteomics (Methods in Molecular Biology).* Humana Press, Totowa, New Jersey. 2007.
[16] E. Mortz, P.B. O'Connor, P. Roepstorff, N.L. Kelleher, T.D. Wood, F.W. McLafferty and M. Mann, "Sequence tag identification of intact proteins by matching tanden mass spectral data against sequence data bases," *Proc. Natl. Acad. Sci. USA*, vol. 93, pp. 8264-8267. 1996.
[17] J. Novák, "Aplikace metrických indexovacích metod na data získaná hmotnostní spektrometrií", *Diploma Thesis*, CTU in Prague, Faculty of Electrical Engineering, Czech Republic. 2008.
[18] J.V. Olsen, S. Ong and M. Mann, "Trypsin Cleaves Exclusively C-terminal to Arginine and Lysine Residues," *Molecular & Cellular Proteomics*, vol. 3, pp. 608-614. 2004.
[19] M. Patella, "Similarity Search in Multimedia Databases", *Ph.D. Thesis*, Dipartmento di Elettronica Informatica e Sistemistica, Bologna, http://www-db.deis.unibo.it/Mtree/index.html. 1999.
[20] G.A. Petsko and D. Ringe, *Protein Structure and Function (Primers in Biology).* New Science Press Ltd, London, UK. 2004.
[21] P.A. Pevzner, Z. Mulyukov, V. Dančík and Ch.L. Tang, "Efficiency of Database Search for Identification of Mutated and Modified Proteins via Mass Spectrometry," *Genome Research*, vol. 11, no. 2, pp. 290-299. 2001.
[22] 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 Oxford Journal*, vol. 22, no. 12, pp. 1524-1531. 2006.
[23] P. Zezula, G. Amato, V. Dohnal and M. Batko, *Similarity Search: The Metric Space Approach (Advances in Database Systems).* Springer, New York, USA. 2006.