Journal of Discrete Algorithms 13 (2012) 19–31

Contents lists available at SciVerse ScienceDirect

Journal of Discrete Algorithms

www.elsevier.com/locate/jda

Non-metric similarity search of tandem mass spectra including posttranslational modifications

Jiří Novák *, Tomáš Skopal, David Hoksza, Jakub Lokoč

SIRET Research Group, Department of Software Engineering, Faculty of Mathematics and Physics, Charles University in Prague, Malostranské nám. 25, 118 00 Prague, Czech Republic

ARTICLE INFO

Article history:
Available online 29 October 2011

Keywords:
Similarity search
Tandem mass spectrometry
Metric access methods
Posttranslational modifications
Peptide identification

ABSTRACT

In biological applications, the tandem mass spectrometry is a widely used method for determining protein and peptide sequences from an “in vitro” sample. The sequences are not determined directly, but they must be interpreted from the mass spectra, which is the output of the mass spectrometer. This work is focused on a similarity-search approach to mass spectra interpretation, where the parameterized Hausdorff distance (d_{HP}) is used as the similarity. In order to provide an efficient similarity search under d_{HP}, the metric access methods and the TriGen algorithm (controlling the metricity of d_{HP}) are employed. Moreover, the search model based on the d_{HP} supports posttranslational modifications (PTMs) in the query mass spectra, what is typically a problem when an indexing approach is used. Our approach can be utilized as a coarse filter by any other database approach for mass spectra interpretation.

1. Introduction

Proteins, organic molecules made of amino acids, are the basis of all living organisms. They are essential for construction of cells and for their proper function [15]. For bioinformatics purposes, a protein can be understood as a linear sequence over 20-letter subset of the English alphabet,[1] where each letter corresponds to an amino acid. A protein sequence must be determined from an “in vitro” protein sample, while tandem mass spectrometry is a very fast and popular method for this task. The proteins in the sample are split by enzymes into shorter pieces called *peptides*, and these are subsequently analyzed by the tandem mass spectrometer [8]. However, instead of direct production of the desired peptide sequences, the spectrometer outputs a set of experimental mass spectra[2] that have to be *interpreted* in order to obtain the peptide sequences. In particular, the interpretation of an experimental spectrum may be accomplished by means of similarity search.

In order to interpret an experimental spectrum, a database D_P of known protein sequences (e.g., MSDB [11]) can be employed. The peptide sequences and their hypothetical spectra are generated from the database D_P, forming a virtual database D_S of mass spectra. Then, the experimental spectrum is used as a query object and the database D_S is searched for its nearest neighbor spectrum (the most similar spectrum from D_S). The experimental spectrum is then interpreted as a peptide sequence corresponding to the spectrum found as the nearest neighbor.

The interpretation of spectra is often complicated by posttranslational modifications (PTMs) occurring in the query. The PTMs are usually not supported in existing similarity approaches among which using of cosine distance is popular.

* Corresponding author.
E-mail address: novak@ksi.mff.cuni.cz (J. Novák).
URL: http://www.siret.cz/novak (J. Novák).

[1] The letters B, J, O, U, X and Z are omitted.

[2] Each spectrum in the set corresponds to one peptide.

1570-8667/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.jda.2011.10.003

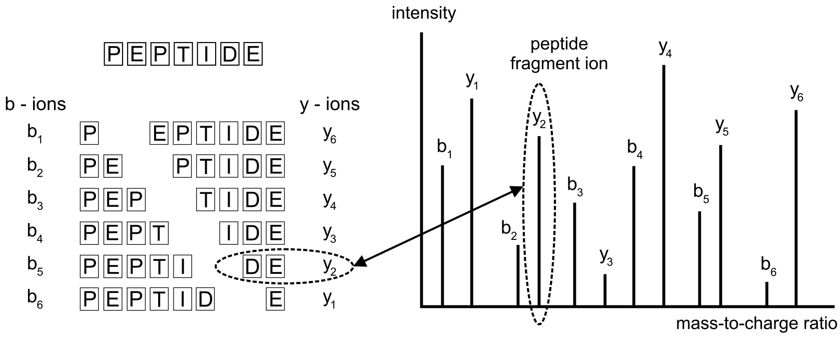

Fig. 1. An example of a mass spectrum.

1.1. Paper contribution

We present the non-metric parameterized Hausdorff distance d_{HP}, which exhibits better correctness of mass spectra interpretation than the cosine similarity does. Moreover, we propose a technique for efficient similarity search in a database of mass spectra indexed under d_{HP}, where for indexing we employ metric access methods (MAMs). In order to use MAMs efficiently, prior to indexing we utilize the TriGen algorithm to control the metricity of d_{HP}. The MAM, which we have chosen in our study, is the M-tree.

Due to the complexity of similarity search of mass spectra with PTMs, this problem is usually neglected in existing indexing approaches. Here, we extend the approach based on d_{HP} to support processing of spectra including PTMs. This extension can be also used in the approaches for mass spectra interpretation based on the cosine similarity.

2. Related work

We briefly describe the structure of data captured by the mass spectrometer and the common techniques employed for mass spectra interpretation using the database search approach.

2.1. Mass spectrometry fundamentals

The mass spectrum is a histogram of peaks corresponding to fragment ions (Fig. 1). A peak is represented by a pair $(\frac{m}{z}, I)$, where $\frac{m}{z}$ is the ratio of mass and charge, and I is the intensity of a fragment ion occurrence. For our purposes it is sufficient to consider $z = 1$ only, thus the ratios $\frac{m}{z}$ are equal to the mass m of fragment ions in Daltons.[3] The precursor mass m_p (the mass of a peptide before splitting) and charge z_p are also provided as an additional information for each peptide spectrum captured by the spectrometer.

In a mass spectrum, there are several types of fragment ions that are highly important for correct peptide sequence identification. The most frequent types of fragment ions with well predictable structure are y-ions and b-ions.[4] Each type of fragment ions forms a ion series, e.g., y-ions series or b-ions series (Fig. 1). The completeness of y-ions and b-ions series is crucial for correct spectra interpretation, because the mass difference between two neighboring peaks in one series, e.g., y_i and y_{i+1} corresponds to a mass of one amino acid.

Often, many of the y-ions or b-ions may never arise in the spectrometer and thus the number of missing y-ions and b-ions is too high to correct mass spectra interpretation. In fact, more than 85% of spectra captured by the spectrometer cannot be interpreted neither by an algorithm nor manually because the split process generates non-standard fragments. However, there are more factors making the interpretation complex. Up to 80% of peaks in each experimental spectrum may correspond to fragment ions with very complicated or unpredictable chemical structure and they make the recognition of y-ions and b-ions difficult. Such peaks are regarded as noise.

2.1.1. Posttranslational modifications

The interpretation of spectra is often complicated due to chemical modifications of amino acids, because masses of amino acids are changed in that case and thus peaks are shifted. This may happen during a sample preparation for the mass analysis, during the mass analysis in the spectrometer or after the translation of proteins in organisms. The last are so-called posttranslational modifications (PTMs; Fig. 2). Since it is not necessary to distinguish the modifications in our study, we use the term PTMs for all the modification types. The database UNIMOD [25] gathers discovered protein modifications for the mass spectrometry. At the time of writing this paper, there were about 660 known modifications.

[3] Dalton (Da) is a unit of the relative atom mass.

[4] In fact, more types of fragment ions with predictable structure may arise in the spectrometer, but many of them occur very rarely.

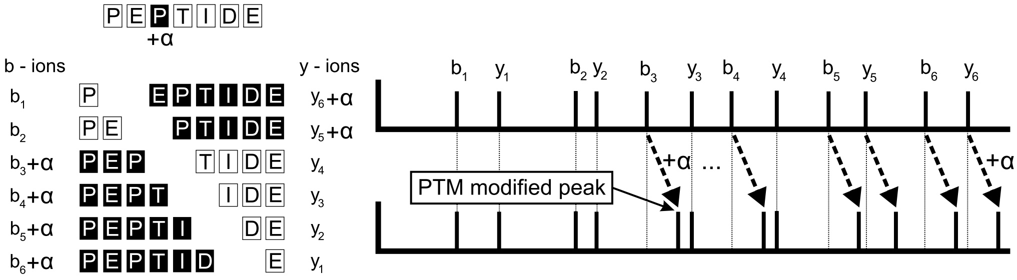

Fig. 2. A peptide with a PTM α (black peptide fragments are affected by the PTM — their masses are modified and corresponding peaks are shifted).

2.2. Similarity search

The best way how the mass spectra may be interpreted is to search a database of already known or predicted peptide (protein, respectively) sequences [8,19]. There are hypothetical mass spectra generated from peptide sequences, and an algorithm (mostly sequential) is used for similarity comparison of an experimental (query) spectrum with the hypothetical (database) spectra. The only difference is that fragment ions intensities cannot be generated from peptide sequences.[5] The basic similarity functions for comparison of an experimental spectrum with the hypothetical spectra generated from the database of protein sequences are, e.g., SPC [7] (shared peak count; in fact, the Hamming distance on boolean vectors, see Fig. 4), spectral alignment [16] (kind of dynamic programming distance on boolean vectors), SEQUEST-like scoring [20]. The most common tools for mass spectra interpretation based on the similarity search in a database are SEQUEST [20], MASCOT [10], ProteinProspector [17], OMSSA [5], etc. A few approaches for interpretation of spectra with PTMs were proposed [9,12,16,23,24].

2.2.1. Metric indexing

Since protein sequence databases grow rapidly and a sequential scan of the whole database becomes slow and inefficient, there is a need for utilization of index structures. A few methods for mass spectra interpretation based on metric access methods were proposed. Metric space approaches are usually based on variants of the cosine similarity (Section 4.1). One of them uses local sensitive hashing to preprocess the database [4], another uses the MVP-tree [18]. The latter approach defines two alternatives of the cosine similarity. The first is called fuzzy cosine distance, while the other is called tandem cosine distance.

In general, a disadvantage of indexing approaches is that they usually do not support the search of spectra with PTMs.

3. Metric access methods

Since our approach to mass spectra interpretation is based on metric similarity search, we need to briefly summarize the main points concerning metric access methods (MAMs) [26] and their applicability. The MAMs were designed for efficient search in databases where a metric distance $d(x, y)$ is employed as the similarity function. The metric distance is a function that satisfies postulates of identity, symmetry, non-negativity and triangle inequality [26]. The metric postulates (especially the triangle inequality) are crucial for MAMs, in order to correctly organize database objects within metric regions and to prune irrelevant regions while searching. The MAMs usually support range and k-NN (k-nearest neighbor) queries. Among the vast number of MAMs developed so far, in our approach we have utilized the M-tree.

3.1. M-tree

The *M-tree* [3] is a dynamic (updatable) index structure that provides good performance in secondary memory, i.e., in database environments. The M-tree index is a hierarchical structure, where some of the data objects are selected as centers (also called local *pivots*) of ball-shaped regions, while the remaining objects are partitioned among the regions in order to build up a balanced and compact hierarchy of data regions. While inner nodes contain *routing entries* associated with metric regions, leaf nodes are represented by *ground entries* containing data objects or identifiers uniquely identifying the data (Fig. 3). When performing a query, the M-tree is traversed from the root, while the subtrees the regions of which overlap the query region must be searched as well, recursively.

3.2. Intrinsic dimensionality

The requirement on metric postulates is crucial for MAMs to index the database, however, the postulates alone do not guarantee an efficient query processing. The efficiency limits of any MAM also heavily depend on the distance distribution

[5] But it does not cause any problems, because intensities are only a secondary information used for a noise filtering from the experimental spectra.

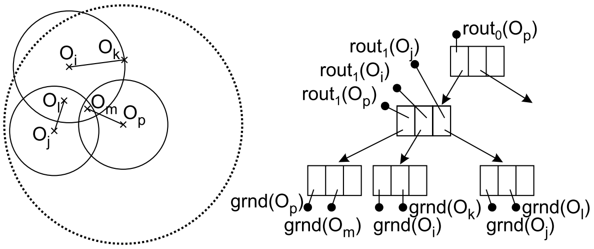

Fig. 3. M-tree.

in the database S, and can be formalized by the concept of *intrinsic dimensionality* $\rho(S,d) = \frac{\mu^2}{2\sigma^2}$, where μ is the mean and the σ^2 is the variance of the distance distribution [2]. In other words, the intrinsic dimensionality is low if the data form tight clusters. Hence, the database can be efficiently searched by a MAM, because a query overlaps only a small number of clusters. On the other hand, a high intrinsic dimensionality (say, $\rho > 10$) indicates most of the data objects are more or less equally far from each other. Hence, in intrinsically high-dimensional database there do not exist clusters, while the search deteriorates to sequential search.

3.3. Non-metric and approximate search

The applicability of MAMs can be extended beyond the metric space model, so that MAMs could be used also for non-metric and/or approximate similarity search. In particular, given a *semi-metric distance* $d(x,y)$ (a metric distance violating the triangle inequality) and a database, the triangle inequality can be added to the semi-metric, so that we obtain a metric modification $f(d(x,y))$ that could be used for similarity search instead. Hence, the MAMs can be correctly used to index and search the database using the metric modification. Moreover, the enforcement of the triangle inequality could be only partial, where the "partial" metric distance could be used for approximate search by MAMs.

3.3.1. TriGen algorithm

The TriGen algorithm [21] was proposed to keep a user-controlled amount of triangle inequality in a semi-metric distance. The idea is based on utilization of a T-modifier, which is either a concave or a convex increasing function f, such that $f(0) = 0$. A concave function f, when applied on a semi-metric, increases the number of triplets $\langle f(d(x,y)), f(d(y,z)), f(d(x,z))\rangle$ that form a triangle (so-called *triangle triplets*), and so improves the triangle inequality fulfillment of $f(d)$. On the other hand, a convex T-modifier f does the opposite — it decreases the number of triangle triplets. Simultaneously, a concave modification $f(d)$ increases the intrinsic dimensionality, as it inhibits the differences between distances. Conversely, a convex modification $f(d)$ decreases the intrinsic dimensionality, as it magnifies the differences between distances. Formally, the proportion of triplets that are NOT triangular in a sample of examined triplets is called the *T-error*. Given a user-defined T-error tolerance, the TriGen algorithm was designed to find a T-modifier for which the intrinsic dimensionality $\rho(S, f(d))$ is minimized, while the T-error does not exceed the tolerance.

In order to automate the search for the optimal T-modifier, the TriGen works with so-called *T-bases* $f(v,w)$. A T-base is a T-modifier with an additional parameter w, that aims to control to convexity or concavity of f. For $w > 0$, the f gets more concave, for $w < 0$ it gets more convex, and for $w = 0$ we get the identity $f(v,0) = v$. A simple T-base used by TriGen is the Fractional-Power base (FP-base) (Eq. (1)).

$$\mathrm{FP}(v,w) = \begin{cases} v^{\frac{1}{1+w}} & \text{for } w > 0 \\ v^{1-w} & \text{for } w \leqslant 0 \end{cases} \tag{1}$$

The modified distance $f(d)$ determined by TriGen can be then employed by any MAM for an exact but slower (T-error tolerance is zero, so ρ gets higher) or only an approximate but fast (T-error tolerance is positive, so ρ gets smaller) similarity search (metric or non-metric).

4. Similarity functions employed in mass spectra interpretation

Although the TriGen algorithm (Section 3.3.1) allows to use MAMs also with non-metric distances, it does not guarantee that a particular non-metric distance modified into metric will be suitable for indexing by MAMs. In particular, a highly non-metric distance (exhibiting high T-error) is modified by TriGen very aggressively to achieve zero T-error, which means the resulting metric will imply high intrinsic dimensionality of the database, thus making it not indexable. Hence, when designing a new similarity that should be indexable by MAMs, the attention must be given not only to the semantics of the similarity/effectiveness, but also to its indexability/efficiency (low both, the T-error and intrinsic dimensionality).

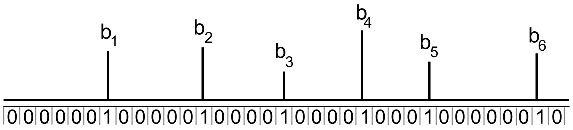

Fig. 4. A high-dimensional boolean representation of a mass spectrum.

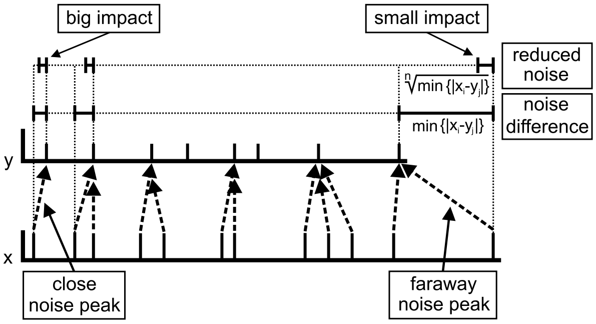

Fig. 5. The fundamentals of d_{HP} (the dashed arrows indicates the closest peaks in $\vec{y}$ to the peaks in $\vec{x}$).

4.1. Cosine similarity

The cosine similarity and its metric form, the angle distance, are commonly mentioned in mass spectrometry literature for peptide mass spectra interpretation [1,4,9,18,22]. The cosine similarity requires a representation of mass spectra as high-dimensional boolean vectors (Fig. 4). For example, let the range of $\frac{m}{z}$ values in a mass spectrum be 0–2000 Da and let it be divided in subintervals of 0.1 Da. The mass spectrum is then represented by a 20 000-dimensional boolean feature vector having ones at places corresponding to intervals for which the $\frac{m}{z}$ value is nonzero in the spectrum. Instead of storing the high-dimensional sparse vector x, there is usually a compact representation $\vec{x}$ used, where the positions of ones in x (i.e., dimensions in which the values of x are nonzero) are substituted with values of the compact vector $\vec{x}$. The compact representation of vector x in Fig. 4 is $\vec{x} = \langle 7, 13, 18, 23, 27, 34\rangle$. We use a semi-metric variant d_A of the angle distance (Eq. (4)) based on the compact representation, where $\dim(\vec{x})$ is the length/dimension of $\vec{x}$ (the number of peaks/ones) and ξ is a mass error tolerance.

$$d_a(\vec{x}_i, \vec{y}_j) = \begin{cases} 0, & \text{if } |\vec{x}_i - \vec{y}_j| > \xi \\ 1, & \text{else} \end{cases} \tag{2}$$

$$a(\vec{x}, \vec{y}) = \frac{\sum_{x_i \in \vec{x}} \max_{y_j \in \vec{y}} \{d_a(\vec{x}_i, \vec{y}_j)\}}{\sqrt{\dim(\vec{x}) \dim(\vec{y})}} \tag{3}$$

$$d_A(\vec{x}, \vec{y}) = \arccos\bigl(a(\vec{x}, \vec{y})\bigr) \tag{4}$$

4.2. Parameterized Hausdorff distance

The parameterized Hausdorff distance d_{HP} (Eq. (7)), suitable for the similarity search in protein sequence-derived databases of theoretical peptide mass spectra, was proposed in [13]. $\vec{x}$ and $\vec{y}$ represent vectors of $\frac{m}{z}$ ratios and $\dim(\vec{x})$ is the length of the vector $\vec{x}$. The internal distance d_h measures the difference between two values, while only distances exceeding threshold ξ (mass error tolerance) are taken into account.

$$d_h(\vec{x}_i, \vec{y}_j) = \max\bigl(0, |\vec{x}_i - \vec{y}_j| - \xi\bigr) \tag{5}$$

$$h(\vec{x}, \vec{y}) = \frac{\sum_{\vec{x}_i \in \vec{x}} \sqrt[n]{(\min_{\vec{y}_j \in \vec{y}} \{d_h(\vec{x}_i, \vec{y}_j)\})}}{\dim(\vec{x})} \tag{6}$$

$$d_{HP}(\vec{x}, \vec{y}) = \max\bigl(h(\vec{x}, \vec{y}), h(\vec{y}, \vec{x})\bigr) \tag{7}$$

d_{HP} is a semi-metric and it reduces the impact of noise peaks using nth root (Fig. 5). The d_{HP} works as follows. First, the value/peak in the minimal distance in the vector/spectrum $\vec{y}$ is found for each peak in $\vec{x}$. The nth root is applied on each of the minimal distances and the sum is computed. The nth root causes that pairs of noise peaks in small distances (exceeding a small error tolerance ξ) have big contributions in the sum and vice versa pairs of noise peaks in big distances have small contributions in the sum (in order to decrease their impact on the sum). Since the number of peaks in compared spectra

Table 1
Intrinsic dimensionality ρ and empirically determined FP(v,w) modifiers for d_{HP} ($n = 30$) and d_A.

T-error	d_{HP}		d_A	
	ρ	w	ρ	w
0	88.5	−0.17	158.1	−0.84
0.01	5.2	−4.44	11.1	−7.43
0.02	4.0	−5.23	8.5	−8.94
0.03	3.5	−5.71	7.1	−10.01
0.04	3.2	−6.08	6.3	−10.92
0.05	2.9	−6.40	5.7	−11.65
0.06	2.8	−6.64	5.2	−12.34
0.07	2.6	−6.87	4.8	−13.00
0.08	2.5	−7.06	4.5	−13.63
0.09	2.4	−7.25	4.2	−14.28
0.1	2.3	−7.42	3.9	−14.92

may be different, an average is computed. The process is repeated with $\vec{x}$ and $\vec{y}$ vectors switched and maximum value is selected to obtain a symmetric measure.

Since the values in $\vec{x}$ and $\vec{y}$ are ordered, the d_{HP} computation is of linear complexity [13] (unlike the general Hausdorff distance [26]). Moreover, using of the time expensive nth root function does not cause any problem, because the range of mass corresponding to generated peptide sequences is limited and thus a table of the roots can be precomputed. It was shown that interpretation using d_{HP} exhibits better efficiency and effectiveness than cosine similarity commonly mentioned in mass spectrometry literature [13].

4.3. *TriGen-based modification*

d_{HP} and d_A are semi-metric distances, the T-error for each of them is very low (below 0.001) but the intrinsic dimensionality is very high (above 88 for d_{HP} and above 158 for d_A). Thus, we used TriGen to improve the intrinsic dimensionality, setting the T-error tolerances to be 0–0.1. Note that d_{HP} and d_A must be normalized to $\langle 0, 1\rangle$ in order to employ the TriGen. The d_{HP} is normalized by $\sqrt[n]{d_h^{\max}}$, where $d_h^{\max}$ is the maximal possible value in a compact vector (i.e., the dimension of the high-dimensional representation). The d_A is normalized by $\frac{\pi}{2}$.

For all the T-error tolerances, the TriGen found convex T-modifiers ($w < 0$), so the intrinsic dimensionality was reduced (down to 2 for T-error tolerance 0.1). The resulting modifiers determined by TriGen for d_{HP} ($n = 30$) and d_A are shown in Table 1.

5. Interpretation using similarity search

The entire process of peptide mass spectra interpretation we propose, incorporating the previously defined measures, can be summarized as follows:

Indexing

1. Each protein sequence in the database is split to shorter peptide sequences. The rules for the splitting are determined by an enzyme. The most common enzyme is trypsin, which splits the protein chains after each amino acid K (lysine) and R (arginine) if they are not followed by P (proline) [14]. However, even if the splitting sites are well predictable, the process is not perfect in practice and some missed cleavage sites can occur. The maximum number of missed cleavage sites $\max_{cs}$ is adjusted as a parameter.
2. The $\frac{m}{z}$ values of y- and b-ions are generated in ascending order for each peptide sequence, while each sequence corresponds to one indexed vector. The vector for the peptide sequence of the length l has the dimension $2(l-1)$, see Fig. 1.
3. The vectors are indexed by a MAM (e.g., by the M-tree) under d_{HP} or d_A modified by the TriGen (Section 3.3.1).

Querying/Interpretation

1. The experimental spectrum is preprocessed before the interpretation. The p peaks with highest intensity I from the experimental spectrum are selected and they form a query corresponding to a vector of their $\frac{m}{z}$ values.
2. A kNN query is processed by the MAM, while the correct peptide sequence corresponding to the spectrum is obtained as the first neighbor in many cases. However, in real-world applications we need to provide more nearest neighbors, because an additional scoring algorithm could select a different peptide as the correct one from the kNN set. Such refining algorithm could be, e.g., SPC, spectral alignment, SEQUEST-like scoring (Section 2.2).

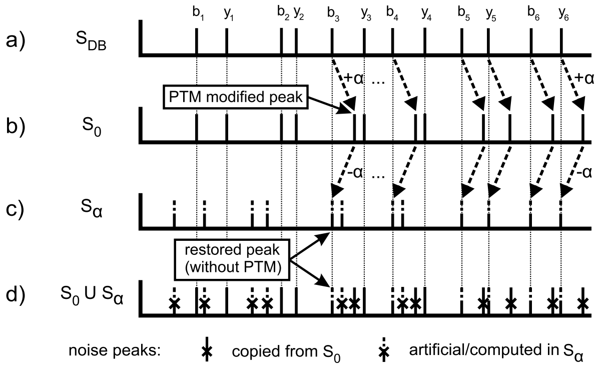

Fig. 6. Dealing with PTMs (S_{DB} corresponds to S_0 with PTM α happened at position 3 in the respective peptide sequence).

In the experimental section we assume that a mass spectrum is successfully interpreted if the correct peptide sequence is among the k nearest neighbors (regardless of its position in the kNN result). Such an approach is often employed and the scoring is then handled separately. Hence, the overall setup of our method can be utilized as a coarse filter by any other database approach for mass spectra interpretation.

5.1. Dealing with PTMs

If an experimental spectrum contains PTMs, some peaks in the spectrum are shifted. The shifts depend on the positions of PTMs in the peptide, i.e., which amino acids in the sequence have modified mass (Fig. 2). There are two basic ways to support identification of the spectra with PTMs. First, all peaks in the database-generated spectra can be shifted for any PTM (or any combination of PTMs) and indexed by a MAM, while the query is unchanged. Since the number of known PTMs is high [25], we use the other way — the modification of the query, while the database remains unchanged. The entire process of the query construction for one random PTM α can be summarized as follows:

1. Let S_{DB} be the database-generated spectrum of a peptide sequence (Fig. 6a). Let $S_0 = \langle m_1, \ldots, m_p\rangle$ (Fig. 6b) be an experimentally taken (i.e., captured by the mass spectrometer) peptide mass spectrum with p peaks (mass-to-charge ratios with $z = 1$).
2. When a PTM α (e.g., $\alpha = 57$) happens at an unknown position i in the peptide, only m_i and some of the following peaks are shifted. Since we cannot predict this position, the entire spectrum is shifted by $-\alpha$. A shift of the spectrum S_0 for the PTM α is a vector $S_\alpha = \langle m_1 - \alpha, \ldots, m_p - \alpha\rangle$ (Fig. 6c). Thus peaks shifted by α in S_0 have their “unshifted” counterparts in S_α.
3. S_0 and S_α are joined (Fig. 6d), where the union of spectra $S_0 \cup S_\alpha$ is a sorted vector of all peaks in the spectra S_0 and S_α.
4. While S_0 forms the query for an unmodified spectrum, the query for the spectrum with PTM α is $S_{\mathrm{I}} = S_0 \cup S_\alpha$.

A disadvantage is that two other types of noise peaks occur in queries. First, the peaks shifted by PTM “in vitro” in S_0, which are superfluous in the union $S_0 \cup S_\alpha$. Second, the artificial noise peaks computed in S_α, which were not modified by PTM and they should not have been shifted in S_α. These two types of noise peaks cannot be removed, because we are not able to recognize them. Since mass spectra contain up to 80% of noise peaks and d_{HP} is able to reduce them, the other noise peaks are reduced as well.

In case of two PTMs α and β, the query is represented by spectrum $S_{\mathrm{II}} = S_0 \cup S_\alpha \cup S_\beta \cup S_{\alpha+\beta}$, where $\alpha + \beta$ are peaks shifted by both modifications at once. In case of three PTMs α, β and γ, the query is represented by spectrum $S_{\mathrm{III}} = S_0 \cup S_\alpha \cup S_\beta \cup S_\gamma \cup S_{\alpha+\beta} \cup S_{\alpha+\gamma} \cup S_{\beta+\gamma} \cup S_{\alpha+\beta+\gamma}$, etc. Since the length of peptide sequences is limited, the number of PTMs per spectrum usually does not exceed 2 or 3 (Table 2). Therefore the maximum number of spectra unified in the query, which might be up to 2^q for q PTMs, is not reached in practice.

A way, how to simplify the query, is to limit the maximum number of simultaneously occurring PTMs n_s per spectrum. For example, if $n_s = 1$ the query spectrum S_{III} is reduced to $S'_{\mathrm{III}} = S_0 \cup S_\alpha \cup S_\beta \cup S_\gamma$. Another example, if $n_s = 2$ and each of PTMs α and β can be repeated up to 2×, we obtain the query spectrum $S_{\mathrm{IV}} = S_0 \cup S_\alpha \cup S_\beta \cup S_{\alpha+\beta} \cup S_{2\times\alpha} \cup S_{2\times\beta}$, etc.

6. Experiments

In the experiments, we used a union of the collections Amethyst and Opal [6] of experimental tandem mass spectra. The collections are formed from the mass spectra of peptides founded in the human genome and they contain mass spectra with PTMs (Table 2). The database used in our experiments is an extension of the list of correct protein sequences assigned

Table 2
The number of PTMs per spectrum and the number of spectra in the collections Amethyst and Opal.

Num. of PTMs	0	1	2	3	4	5	6
Amethyst	1095	371	85	13	2	2	1
Opal	239	237	51	8	1	0	0

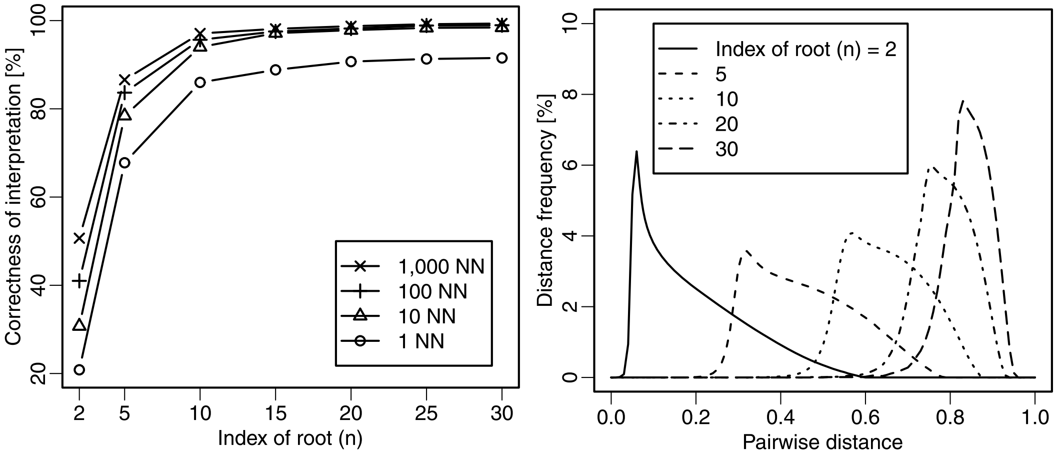

Fig. 7. Correctness of interpretation of d_{HP} — sequential scan (a) and distance distributions (b).

to the experimental mass spectra. The database was extended with protein sequences from MSDB (release 08-31-2006) [11], it contained 100 000 protein sequences (5 612 211 peptide sequences).

In the experiments, we measured two quantities. First, we computed the correctness of mass spectra interpretation (or correctness of peptide sequence identification) as a ratio of correctly assigned peptide sequences to all spectra from a query set. As mentioned in Section 5, we assume that a query spectrum is correctly assigned to the peptide sequence if the correct peptide sequence is among the k nearest neighbors to the query spectrum. Second, we measured the average query time per one mass spectrum interpretation.

All experiments were carried out on a machine with 2 processors Intel Xeon E5450 (8 cores × 3 GHz) with 8 GB RAM and 64-bit OS Windows Server 2008 R2.

The following settings were used unless otherwise specified — the d_{HP} was computed with $n = 30$, the splitting enzyme was trypsin, the maximum missed cleavage sites ($\max_{cs}$) was set to 1, the mass error tolerance (ξ) was 0.4 Da, 50 peaks with the highest intensity were selected from experimental spectra, y- and b-ions were generated to the hypothetical mass spectra.

6.1. Sequential scan

First, d_{HP} was employed with the sequential scan of the whole database of hypothetical mass spectra, while the correctness of interpretation and average query time were measured on the experimental spectra lacking PTMs. The correctness of interpretation was higher with increasing index of the root n (Fig. 7a). The correctness of interpretation was up to 98.3% ($n = 30$; 10 NN queries). The average query time was 14.4 s. The correctness of interpretation for d_A was 95.7% (10 NN queries) and the average query time was 9.8 s.

6.2. Improving the Indexability

A disadvantage of the nth root function in d_{HP} is that intrinsic dimensionality ρ increases with the increasing n, hence the difference between MAMs and sequential scan decreases with increasing n. In Fig. 7b see the distance distributions under d_{HP} (not modified by TriGen) for various n. The x-axis represents normalized distances in the database. The more the distribution is pushed to the right, the higher the intrinsic dimensionality.

In Fig. 8 observe the impact of T-error tolerance on the distance distributions obtained using the TriGen-modified d_{HP} and d_A considering the FP-base. Obviously, higher T-error tolerance leads to more convex T-modifier, hence to lower intrinsic dimensionality (distance distributions pushed to the left).

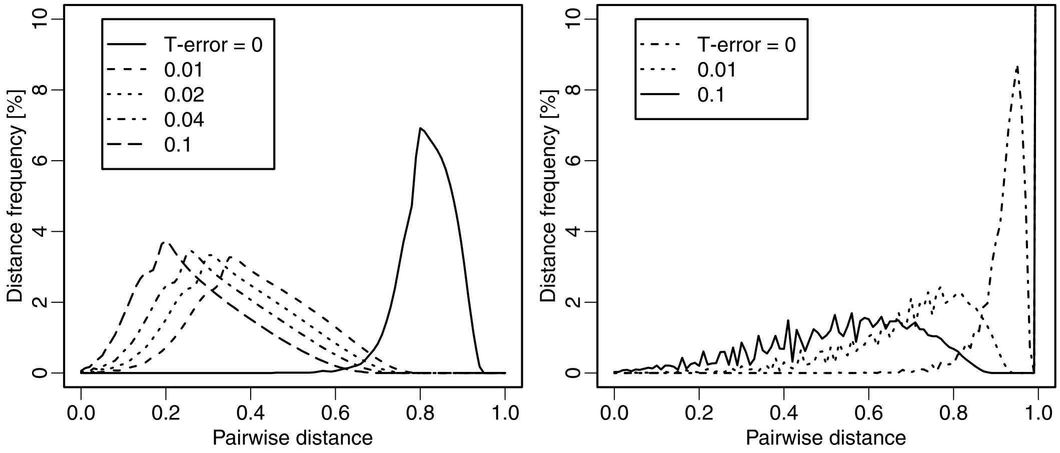

Fig. 8. d_{HP} (a) and d_A (b) — distance distributions (modified by TriGen).

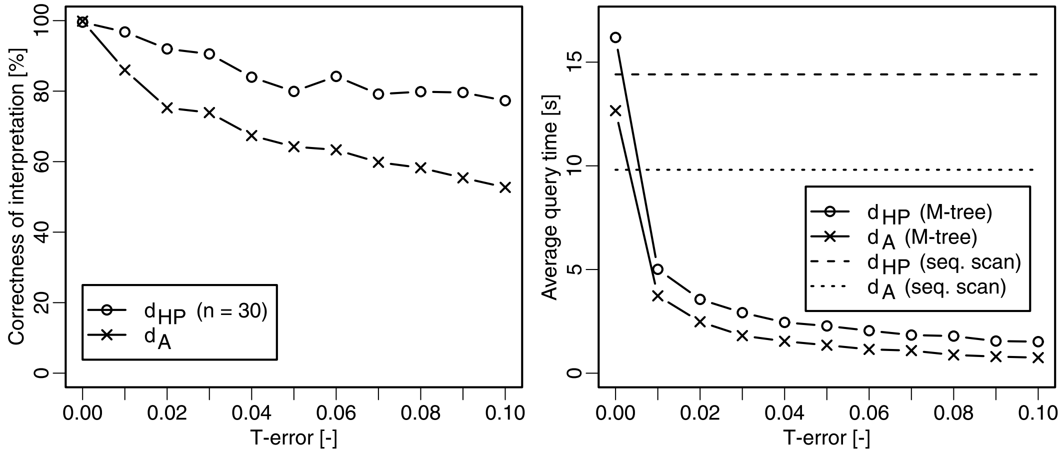

Fig. 9. d_{HP} and d_A — correctness of interpretation (a) and average query time (b) on M-tree.

6.3. Speed-up using M-tree

In order to verify the behavior of d_{HP} and d_A predicted in Section 6.2, we employed 1000 NN queries for various T-error tolerances on the M-tree (Fig. 9). The correctness of interpretation was higher for d_{HP} than for d_A with increasing T-error tolerance. On average, it was 1.3× higher than for d_A. The d_{HP} was 4.9× faster than sequential scan, while the correctness of interpretation was more than 90% (T-error = 0.03). The d_A was 5.4× faster at the same T-error tolerance, but correctness was only 73.9%. The average query time was 14.4 s for d_{HP} and 9.8 s for d_A on the sequential scan (Section 6.1).

6.4. Searching with PTMs

In this section we show that our proposed measures are also capable of interpretation of mass spectra containing PTMs.

6.4.1. Utilization of d_{HP}

The correctness of interpretation was taken using kNN queries without/with (Table 3) the support of the search of PTMs using the sequential scan. We tested 467 spectra containing one PTM, 77 spectra containing two PTMs and 10 spectra containing three PTMs. The following PTMs $\alpha \in \{-28, -17, -14, 1, 14, 16, 28, 57\}$, pairs of PTMs $\{\alpha, \beta\} \in \{\{-17, 57\}, \{57, 57\}\}$ and triplets of PTMs $\{\alpha, \beta, \gamma\} \in \{\{-17, 57, 57\}, \{57, 57, 57\}\}$ were searched. The queries S_{I} were performed for spectra with one PTM, S_{II} for spectra with pairs of PTMs and S_{III} for spectra with triplets of PTMs (Section 5.1).

Since modifications might not affect all peaks in the experimental spectrum, the d_{HP} is still partially able to determine the correct peptide sequence without PTMs support. The correctness of interpretation was more than 90% in all cases when PTMs were supported (1000 NN queries). It decreases with increasing number of PTMs per spectrum when smaller kNN queries are used.

Table 3
Correctness of interpretation without/with the support of PTMs in the query mass spectra.

PTMs per spectrum	Correctness of interpretation [%]			
	1 NN	10 NN	100 N N	1000 NN
without the support of PTMs				
1	20.0	41.0	61.7	75.0
2	9.3	18.6	28.7	65.8
3	0	0	0	0
with the support of PTMs				
1	69.9	84.0	94.3	98.9
2	24.8	55.1	76.8	90.8
3	31.0	54.8	61.9	100.0

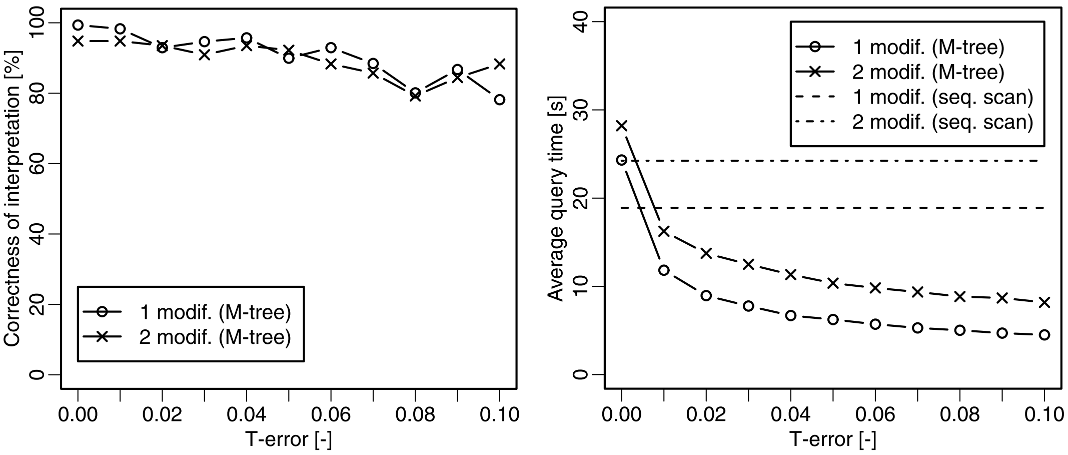

Fig. 10. PTMs — correctness of interpretation (a) and average query time (b) on the M-tree.

The number of peaks in the query increases with increasing number of PTMs to be supported and the average query time increases too. The average query time for the spectra with one PTM was 18.9 s, for the spectra with two PTMs was 24.2 s and for the spectra with three PTMs was 35.6 s.

6.4.2. Queries on M-tree

We performed a set of 1000 NN queries for different T-error tolerances, while the M-tree and d_{HP} were employed. The results for spectra with one and two PTMs are shown in Fig. 10. The M-tree was 3.3× faster for spectra with one PTM and 2.5× faster for spectra with two PTMs than the sequential scan (T-error $= 0.06$), while the correctness of interpretation was still about 90%.

6.4.3. Impact of PTMs setup

The user of a real-life application, who wants to interpret the mass spectra, is usually able to predict some of the PTMs, which may occur in the data-set captured by the spectrometer. Thus the PTMs, which are taken into account during the search, are commonly selected by the user before the search. We are interested in the impact of the query construction (i.e., in the user's well/badly formed choice of PTMs) on the correctness of interpretation.

The correctness of interpretation was taken for the mixture of experimental spectra with modifications $\{0, 57, -17, 16\}$ (none or one modification per spectrum is assumed, i.e., $n_s = 1$). The results were taken for d_{HP} (Table 4) and d_A (Table 5), while the sequential scan was employed for different k in kNN queries. The queries were gradually expanded to cover all the PTMs in the mixture. The PTM supported by the current query extension is indicated by $\downarrow$ (e.g., if the query is changed from S_0 to $S_0 \cup S_{57}$, the PTM $+57$ Da is indicated) and the results for the distance with higher correctness are highlighted.

d_{HP} had better correctness of interpretation in comparison to d_A when smaller kNN queries (1 NN to 100 NN) were used. When 1000 NN queries were employed, d_A had a little bit better correctness than d_{HP} in half the cases. The d_{HP} would be better than d_A, if a bigger protein sequence database and 1000 NN query were used, because false hits in the kNN query worse the correctness of d_A.

Some spectra with PTMs were correctly assigned to the peptide sequences even if the queries were not modified to support them, while the results were noticeably better for d_{HP} than for d_A. The query expansion to cover more PTMs considerably increases the correctness for the spectra with these PTMs and slightly decreases the correctness for the spectra with PTMs, which were covered before the query expansion. In another words, if the user's selection of PTMs to be searched

Table 4
d_{HP} — correctness of interpretation [%].

PTM [Da]	0	57	−17	16	Total
Num. of spectra	1334	280	29	34	1677
S_0	↓				
1 NN	**91.4**	**30.7**	**27.6**	**50.0**	**79.3**
10 NN	**98.3**	**46.1**	**62.1**	**79.4**	**88.6**
100 NN	98.8	**61.4**	**93.1**	**97.1**	**92.4**
1000 NN	99.1	**76.8**	**93.1**	**100.0**	**95.3**
$S_0 \cup S_{57}$		↓			
1 NN	**70.5**	**68.9**	**13.8**	8.8	**70.0**
10 NN	**93.1**	**90.0**	**31.3**	**41.2**	**90.5**
100 NN	**97.4**	**97.9**	**58.6**	**64.7**	**96.1**
1000 NN	98.4	98.9	**82.8**	**85.3**	**97.9**
$S_0 \cup S_{57} \cup S_{-17}$			↓		
1 NN	**63.9**	**57.1**	**51.8**	8.8	**61.5**
10 NN	**84.3**	**80.7**	62.1	**26.5**	**82.1**
100 NN	**93.4**	**93.2**	65.5	**52.9**	**92.1**
1000 NN	95.7	**98.6**	72.4	**79.4**	95.5
$S_0 \cup S_{57} \cup S_{-17} \cup S_{16}$				↓	
1 NN	**46.5**	38.2	**48.3**	**26.5**	**44.7**
10 NN	**71.1**	**65.4**	58.6	**61.2**	**69.7**
100 NN	**84.6**	**83.6**	65.5	**91.2**	**84.2**
1000 NN	92.0	92.9	72.4	**100.0**	91.9

Table 5
d_A — correctness of interpretation [%].

PTM [Da]	0	57	−17	16	Total
Num. of spectra	1334	280	29	34	1677
S_0	↓				
1 NN	84.8	21.8	13.8	17.6	71.7
10 NN	95.7	35.7	37.9	38.2	83.5
100 NN	**99.0**	46.1	58.6	64.7	88.7
1000 NN	**99.6**	63.6	82.3	97.1	93.3
$S_0 \cup S_{57}$		↓			
1 NN	59.4	60.7	3.4	8.8	57.7
10 NN	79.4	79.3	10.3	14.7	76.9
100 NN	92.7	91.8	37.9	38.2	91.0
1000 NN	**99.0**	**99.3**	65.5	55.9	97.6
$S_0 \cup S_{57} \cup S_{-17}$			↓		
1 NN	50.7	54.3	44.8	8.8	50.3
10 NN	70.9	71.4	**72.4**	17.6	69.9
100 NN	88.0	88.6	**89.7**	32.4	87.0
1000 NN	**97.2**	97.1	**100.0**	55.9	**96.4**
$S_0 \cup S_{57} \cup S_{-17} \cup S_{16}$				↓	
1 NN	38.6	**43.2**	17.2	23.5	38.7
10 NN	57.4	62.1	58.6	58.8	58.3
100 NN	77.5	81.1	**79.3**	73.5	78.1
1000 NN	**93.5**	**96.1**	**100.0**	88.2	**93.9**

is too vigorous (i.e., many unnecessary PTMs are selected), the correctness of interpretation might dramatically decrease. On the other hand, if the user omits some of the PTMs, which are presented in the experimental mass spectra, the spectra can be still successfully interpreted.

6.4.4. Search of spectra with more PTMs

The spectra with at most two PTMs ($n_s = 2$) per spectrum were interpreted with d_{HP}, while more complex queries S_{IV} (Section 5.1) with $\alpha = 57$ and $\beta = -17$ were performed using the sequential scan (100 and 1000 NN queries were used). Otherwise stated, we were trying to simulate a more real-life search situation, where two PTMs were given on the input of an application and the query spectra could contain up to both PTMs or did not have to contain any PTM. The results are

Table 6
Search of spectra with more PTMs — correctness of interpretation [%].

PTMs in spectra [Da]	Num. of spectra	Query			
		S_0	S_{I}	S_{II}	S_{IV}
100 NN					
none	1334	98.8	–	–	65.1
+57	280	61.4	97.9	–	66.8
−17	29	93.1	86.2	–	58.6
+57 + 57	64	34.4	–	84.4	65.6
−17 + 57	13	23.1	–	69.2	53.8
1000 NN					
none	1334	99.1	–	–	**82.5**
+57	280	76.8	98.9	–	**86.4**
−17	29	93.1	93.1	–	**68.7**
+57 + 57	64	46.9	–	96.9	**85.9**
−17 + 57	13	84.6	–	84.6	**69.2**

summarized in Table 6. The correctness of interpretation was more than 82% for spectra without PTMs, more than 85% for spectra with up to two PTMs +57 Da and almost 70% for spectra containing one PTM −17 Da or the combination of PTMs −17 Da and +57 Da (1000 NN queries are assumed). We employed the sequential scan with the average query time 32.8 s.

For a comparison, we performed queries S_0 (i.e., PTMs were not supported) and also queries S_{I}, S_{II} (i.e., it was known what PTMs should be found). For the spectra with PTMs, the correctness was in many cases better for S_{IV} than for S_0 but a little bit lower than for S_{I} and S_{II}.

7. Conclusions

The best way how to interpret the tandem mass spectra of peptides is to search a database of already known or predicted protein sequences. We have shown that M-tree and parameterized Hausdorff distance (d_{HP}) is a powerful combination for this task. The d_{HP} models the similarity among the spectra very well and it can be utilized by MAMs when TriGen algorithm is employed. In general, if the T-error is higher, the indexability of the d_{HP} by MAMs is better, the search is faster and the correctness of interpretation is a little bit lower.

Moreover, we have proposed an extension of the d_{HP} approach for the spectra containing posttranslational modifications (PTMs), which are in practice a relatively frequent phenomenon but often neglected in the existing indexing approaches. Since the extension is independent of d_{HP} and MAMs, it can be implemented by other approaches to increase the effectiveness for spectra contaminated by PTMs, e.g., that one based on the cosine similarity.

Acknowledgements

This work has been supported by Czech Science Foundation (GAČR) projects 201/09/0683, 202/11/0968, 201/09/H057 and by the Grant Agency of Charles University (GAUK) project No. 430711.

References

[1] Z.B. Alfassi, On the normalization of a mass spectrum for comparison of two spectra, Journal of the American Society for Mass Spectrometry 15 (3) (2004) 385–387.
[2] E. Chávez, G. Navarro, A probabilistic spell for the curse of dimensionality, in: ALENEX'01, in: LNCS, vol. 2153, Springer, 2001, pp. 147–160.
[3] P. Ciaccia, M. Patella, P. Zezula, M-tree: An efficient access method for similarity search in metric spaces, in: Proc. of 23rd Int. Conf. on VLDB, 1997, pp. 426–435.
[4] D. Dutta, T. Chen, Speeding up tandem mass spectrometry database search: Metric embeddings and fast near neighbor search, Bioinformatics Oxford Journal 23 (5) (2007) 612–618.
[5] L.Y. Geer, et al., Open mass spectrometry search algorithm, Journal of Proteome Research 3 (2004) 958–964.
[6] GPM, the Global Proteome Machine Organization, http://www.thegpm.org/quartz/.
[7] N.C. Jones, P.A. Pevzner, An Introduction to Bioinformatics Algorithms, MIT Press, Cambridge, MA, 2004.
[8] M. Kinter, N.E. Sherman, Protein Sequencing and Identification Using Tandem Mass Spectrometry, John Wiley & Sons, New York, USA, 2000.
[9] J. Liu, et al., Methods for peptide identification by spectral comparison, Proteome Science 5 (3) (2007).
[10] MASCOT, http://www.matrixscience.com/.
[11] Mass Spectrometry Protein Sequence Database, http://www.proteomics.leeds.ac.uk/bioinf/msdb.html.
[12] K. Ning, H. Ng, H. Leong, An accurate and efficient algorithm for peptide and PTM identification by tandem mass spectrometry, Genome Informatics 19 (2007) 119–130.
[13] J. Novák, D. Hoksza, Parametrised Hausdorff distance as a non-metric similarity model for tandem mass spectrometry, in: CEUR Proc. DATESO, 2010, pp. 1–12.
[14] J.V. Olsen, S. Ong, M. Mann, Trypsin cleaves exclusively C-terminal to arginine and lysine residues, Molecular and Cellular Proteomics 3 (2004) 608–614.
[15] G. Petsko, D. Ringe, Protein Structure and Function (Primers in Biology), New Science Press Ltd., London, UK, 2004.
[16] P.A. Pevzner, Z. Mulyukov, V. Dančík, C.L. Tang, Efficiency of database search for identification of mutated and modified proteins via mass spectrometry, Genome Research 11 (2) (2001) 290–299.

[17] ProteinProspector, http://prospector.ucsf.edu/.
[18] S.R. Ramakrishnan, et al., A fast coarse filtering method for peptide identification by mass spectrometry, Bioinformatics 22 (12) (2006) 1524–1531.
[19] R.G. Sadygov, et al., Large-scale database searching using tandem mass spectra: Looking up the answer in the back of the book, Nature Methods 1 (3) (2004) 195–202.
[20] SEQUEST, http://fields.scripps.edu/sequest/.
[21] T. Skopal, Unified framework for fast exact and approximate search in dissimilarity spaces, ACM Transactions on Database Systems 32 (4) (2007) 29.
[22] D.L. Tabb, et al., Similarity among tandem mass spectra from proteomic experiments: Detection, significance and utility, Anal. Chem. 75 (10) (2003) 2470–2477.
[23] S. Tanner, H. Shu, A. Frank, L.-C. Wang, E. Zandi, M. Mumby, P.A. Pevzner, V. Bafna, InsPecT: Identification of posttranslationally modified peptides from tandem mass spectra, Analytical Chemistry 77 (14) (2005) 4626–4639.
[24] D. Tsur, et al., Identification of post-translational modifications by blind search of mass spectra, Nature Biotechnology 23 (2005) 1562–1567.
[25] UNIMOD, http://www.unimod.org/.
[26] P. Zezula, G. Amato, V. Dohnal, M. Batko, Similarity Search: The Metric Space Approach (Advances in Database Systems), Springer, USA, 2006.