Improving the Similarity Search of Tandem Mass Spectra using Metric Access Methods

Jiří Novák, Tomáš Skopal, David Hoksza and 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
http://siret.ms.mff.cuni.cz

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 parametrized 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. We show that similarity search using d_{HP} exhibits better correctness of peptide mass spectra interpretation than the cosine similarity commonly mentioned in mass spectrometry literature. Moreover, the search model using the d_{HP} distance could be extended to support chemical modifications in the query mass spectra, which is typically a problem when the cosine similarity is used. Our approach can be utilized as a coarse filter by any other database approach for mass spectra interpretation.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*indexing methods*; J.3 [**Life and Medical Sciences**]: Biology and Genetics

General Terms

Design, Performance

1. INTRODUCTION

Proteins – organic molecules made of amino acids – are the basis of all living organisms. The proteins are essential for construction of cells and for their proper functioning [22]. For bioinformatic purposes (i.e., in computerized biology), 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* as peptide sequences some other way. 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 P of known protein sequences (e.g., MSDB [17]) could be employed. The peptide sequences and their hypothetical spectra are generated from the database P, forming a database S of mass spectra. Then, the experimental spectrum is used as a query object q and the database S is searched for the nearest neighbor spectrum of q (the most similar spectrum from S). The experimental spectrum is then interpreted as a peptide sequence corresponding to the spectrum found as the nearest neighbor. As functions used to evaluate the similarity between two spectra, variations of the cosine similarity are popular.

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 search in a database of mass spectra indexed under d_{HP}, where for indexing we employ the metric access methods (MAMs). Since d_{HP} is a non-metric distance, the MAMs cannot be used directly, so prior to indexing we utilize the TriGen algorithm to control the metricity of d_{HP}. Among the number of MAMs, we have chosen the M-tree and the Pivot tables in our study. We also show that utilization of cosine similarity for the task of peptide mass spectra interpretation using MAMs is limited.

2. RELATED WORK

We briefly describe the structure of data captured by the mass spectrometer and two basic ways commonly used for mass spectra interpretation. The spectra may be interpreted directly using graph algorithms or by the search in a database of protein sequences.

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

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

SISAP '10, September 18-19, 2010, Istanbul, Turkey.

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 $\left(\frac{m}{z}, I\right)$, where $\frac{m}{z}$ is the ratio of mass and charge, and I is the intensity of a fragment ion occurrence. The charge is supposed for our purposes $z = 1$, so the ratios $\frac{m}{z}$ are equal to the mass m of fragment ions in Daltons[3]. The precursor mass m_p (mass of peptide before splitting) and charge z_p are also provided as an additional information for each peptide spectrum captured by the spectrometer.

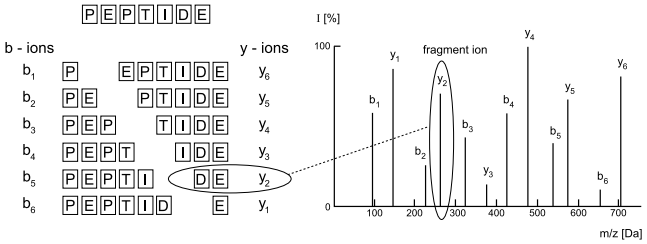

Figure 1: An example of a mass spectrum.

In a mass spectrum, there are several types of fragment ions that are the most important for a correct interpretation. 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 an 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. However, there are more factors making the interpretation complex. Up to 80% of peaks in an experimental spectrum may correspond to fragment ions with very complicated or unpredictable chemical structure and they complicate the recognition of y-ions and b-ions. Such peaks are regarded as noise.

The interpretation of spectra may be also complicated due to chemical modifications of amino acids, because mass of amino acids are changed in that case and thus peaks are shifted. This may happen, e.g., during a sample preparation for mass analysis or in the spectrometer. The database UNIMOD [31] gathers discovered protein modifications for mass spectrometry. At the time of writing this paper, there were about 650 known modifications.

2.2 Graph-based Approaches

The mass spectra may be interpreted directly using graph algorithms (without any reference database). Such approaches are called *de novo* peptide sequencing [4] and they are based on detection of y-ions and b-ions series. A graph is constructed from an experimental spectrum, where a node corresponds to a peak (its mass m) and an edge is ranked with the mass difference between two nodes. The graph is traversed, while paths where weights of edges best fit the mass of amino acids are selected. A problem is that many paths and thus many peptide sequences can be assigned to an experimental spectrum, so the correctness of interpretation of such approaches is low (about 30%). This is due to noise, chemical modifications and the fact that some of y-ions or b-ions may never arise in the spectrometer. Some tools for mass spectra interpretation based on the *de novo* approach are, e.g., PEAKS [15], PepNovo [6] and Lutefisk [14].

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

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

2.3 Similarity Search Approaches

The best way how the mass spectra may be interpreted is to search a database of already known or predicted peptide (protein, respectively) sequences [11, 26]. 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. The basic similarity functions for comparison of the experimental spectrum with hypothetical spectra generated from the database of protein sequences are, e.g., SPC [9] (shared peak count; in fact, the Hamming distance on boolean vectors, see Fig. 3), spectral alignment [23] (kind of dynamic programming distance on boolean vectors) [23], SEQUEST-like scoring [27]. The most common tools for mass spectra interpretation based on search in the database are SEQUEST [27], MASCOT [16], ProteinProspector [24], OMSSA [7], etc.

2.3.1 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 (Sec. 4.1). One of them uses locality sensitive hashing to preprocess the database [5], another uses the MVP-tree [25]. The latter approach defines two alternatives of the cosine similarity. The first is called the fuzzy cosine distance, while the other is called the tandem cosine distance.

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) [32] 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 [32]. 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 and Pivot tables.

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. 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 Pivot Tables

A simple but efficient solution to similarity search represent methods called *pivot tables* (or distance matrix methods) [18]. In general, a set of l objects (so-called pivots) is selected from the database, while for every database object a l-dimensional vector of distances to the pivots is created. The vectors belonging to the database objects then form a distance matrix – the pivot table. When performing a kNN query, a distance vector for the query object q is determined the same way as for a database object. Then, the query is processed on the pivot table such that database object vectors which do not belong to the already retrieved kNN candidates are filtered out from further processing.

3.3 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 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.

3.4 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.

3.4.1 TriGen Algorithm

The TriGen algorithm [28] 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.

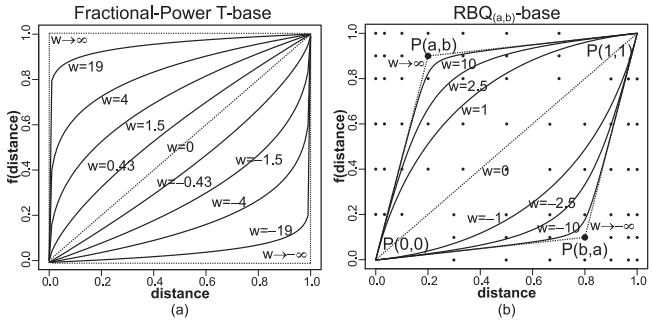

Figure 2: The FP-base and an RBQ$_{(a,b)}$-base.

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) (1), while a more sophisticated T-base is the Rational-Bézier-Quadratic base (RBQ-base) (2), see Fig. 2. Actually, the TriGen uses multiple RBQ-bases – each for a particular pair (a,b).

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

$$\mathrm{RBQ}_{(a,b)}(v,w) = \begin{cases} rbq(v,w,a,b) & \text{for } w > 0 \\ rbq(v,-w,b,a) & \text{for } w \le 0 \end{cases} \quad (2)$$

For description of the *rbq* function, see [28].

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. SIMILARITIES FOR MASS SPECTRA

Although the TriGen algorithm (Sec. 3.4.1) 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.

This could be the case of non-metric distances mentioned in Section 2.3, where the T-error and intrinsic dimensionality was not observed as an important factor for indexability. Hence, when designing a new similarity that has to be indexable by MAMs, the attention should be given not only to the semantics of the similarity, but also to its indexability (low both the T-error and intrinsic dimensionality).

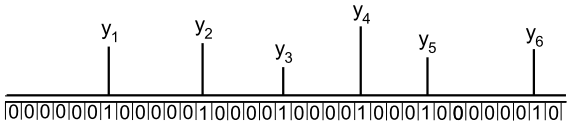

Figure 3: A high-dimensional boolean representation of a mass spectrum.

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 [12, 29, 1, 25, 5]. The cosine similarity requires a representation of mass spectra as high-dimensional boolean vectors (Fig. 3). For example, let the range of $\frac{m}{z}$ values in a mass spectrum be 0-2,000 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 1s 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 1s in x (i.e., dimensions in which the values of x are nonzero) turn into values of the compact vector $\vec{x}$. The compact representation of vector x in Fig. 3 is $\vec{x} = \langle 7, 13, 18, 23, 27, 34\rangle$. We used a semimetric variant d_A of the angle distance based on the compact representation (4), where ξ is a mass error tolerance. Note that subintervals are not bounded as in Fig. 3 because the differences between $\frac{m}{z}$ values are computed.

$$d_a(\vec{x}_i, \vec{y}_j) = \begin{cases} 0, & \text{if } |\vec{x}_i - \vec{y}_j| > \xi \\ 1, & \text{else} \end{cases} \tag{3}$$

$$d_A(\vec{x}, \vec{y}) = \arccos\left(\frac{\sum_{x_i \in \vec{x}} \max_{y_j \in \vec{y}} \{d_a(\vec{x}_i, \vec{y}_j)\}}{\sqrt{\dim(\vec{x})\dim(\vec{y})}}\right) \tag{4}$$

However, the indexability of d_A showed to be inefficient due to the extremely high intrinsic dimensionality. In the experiments (Sec. 5.2) we show that even the utilization of TriGen algorithm with reasonably set T-error tolerance (say, up to 0.2) does not lead to an intrinsic dimensionality acceptable for indexing by MAMs. In particular, many mass spectra are maximally distant (having the angle $\frac{\pi}{2}$, i.e., $d_A = 1$), so that none of the T-modifiers can decrease the intrinsic dimensionality.

An approach for fast indexing under angle distance by MAMs was proposed in [25], where two semimetric alternatives of the cosine similarity have been defined. The first is called the fuzzy cosine distance and it has similar behaviour as d_A. The other is called the tandem cosine distance and it is, in fact, the combination of d_A with the precursor mass filter, i.e., the difference between the precursor mass m_p^x and m_p^y of the compared spectra was joined with d_A in order to reach good indexability (6), where c_1 and c_2 are constants and ξ' is the precursor mass error tolerance. A disadvantage is that the precursor mass filter could limit the capability of managing chemical modifications, because m_p of experimental spectra with modifications can differ by more than a few tens of Daltons from m_p of hypothetical spectra generated from the database of protein sequences.

$$d_{mp}(m_p^x, m_p^y) = \begin{cases} 0, & \text{if } |m_p^x - m_p^y| \le \xi' \\ |m_p^x - m_p^y|, & \text{else} \end{cases} \tag{5}$$

$$d'_A(\vec{x}, \vec{y}) = c_1 d_A(\vec{x}, \vec{y}) + c_2 d_{mp}(m_p^x, m_p^y) \tag{6}$$

4.2 Parametrized Hausdorff Distance

In this paper we employ the parameterized Hausdorff distance, that outperforms the angle distance d_A in both, the indexability by MAMs (i.e., efficiency) and correctness of mass spectra interpretation (i.e., effectiveness).

4.2.1 Logarithmic Distance

The first step towards d_{HP} was the logarithmic distance d_L (8) [19]. The d_L was defined for the compact representations $\vec{x}, \vec{y}$ of the high-dimensional boolean vectors x, y (Sec. 4.1). The logarithmic distance is more robust than the Euclidean distance (L_2). For our application (i.e., mass spectra interpretation) a distance is robust if two vectors $\vec{x}$ and $\vec{y}$ are closer if there are great differences in a small number of their components than if there are small differences in a large number of their components. For example, let us assume vectors $\vec{x} = \langle 200, 300, 400, 500\rangle$, $\vec{y_1} = \langle 200, 300, 460, 500\rangle$ and $\vec{y_2} = \langle 210, 305, 420, 475\rangle$. The missing number 400 in $\vec{y_1}$ with respect to $\vec{x}$ means that the corresponding peak in the mass spectrum is missing. The superfluous number 460 in $\vec{y_1}$ refers to a noise peak, so the vectors $\vec{x}$ and $\vec{y_1}$ should be closer than $\vec{x}$ and $\vec{y_2}$. However, the Euclidean distance (L_2) of the vectors $\vec{x}$ and $\vec{y_1}$ is higher. The d_L distance is lower and thus it models the similarity among mass spectra better ($d_L(\vec{x}, \vec{y_1}) \doteq 1.8$, $d_L(\vec{x}, \vec{y_2}) \doteq 4.4$, $L_2(\vec{x}, \vec{y_1}) = 60$, $L_2(\vec{x}, \vec{y_2}) \doteq 33.9$).

$$d_l(\vec{x}_i, \vec{y}_i) = \begin{cases} \log|\vec{x}_i - \vec{y}_i|, & \text{if } |\vec{x}_i - \vec{y}_i| > 1 \\ 0, & \text{else} \end{cases} \tag{7}$$

$$d_L(\vec{x}, \vec{y}) = \sum_i d_l(\vec{x}_i, \vec{y}_i) \tag{8}$$

4.2.2 Hausdorff distance

In general, the Hausdorff distance d_H [32] is a metric distance defined on sets A and B (10). The d_H is computed such that nearest neighbors from objects in one set are determined to all objects in the other set (both directions), while the maximal distance is reported as the result of d_H. The internal distance d_x operating on the objects of A, B could be any other distance (e.g., L_2 in case of point sets). If the d_x is a metric then the d_H is metric, too.

$$h(A, B) = \max_{a_i \in A}\left\{\min_{b_j \in B}\{d_x(a, b)\}\right\} \tag{9}$$

$$d_H(A, B) = \max(h(A, B), h(B, A)) \tag{10}$$

4.2.3 Parametrized Hausdorff Distance

The parametrized Hausdorff distance d_{HP} (13) is a semimetric, which combines advantages of the d_L and d_H [20]. The d_{HP} uses n^{th} root function instead of logarithm because of higher correctness of interpretation. Also, compact vectors of different dimensions can be used which is

desirable because the mass spectra have different numbers of peaks and thus the compared compact vectors have different lengths. A property inherited from d_H is its robustness – unlike d_L, the values in $\vec{x}, \vec{y}$ are compared based on their closeness, not the same position (dimension) in the vectors. Since the values in $\vec{x}$, $\vec{y}$ are ordered, the d_{HP} computation is of linear complexity (unlike the general d_H) [20]. Moreover, using of the time expensive n^{th} 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.

$$d_e(\vec{x}_i, \vec{y}_j) = \begin{cases} |\vec{x}_i - \vec{y}_j|, & \text{if } |\vec{x}_i - \vec{y}_j| > \xi \\ 0, & \text{else} \end{cases} \quad (11)$$

$$h'(\vec{x}, \vec{y}) = \frac{\sum_{\vec{x}_i \in \vec{x}} \sqrt[n]{\left(\min_{\vec{y}_j \in \vec{y}} \{d_e(\vec{x}_i, \vec{y}_j)\}\right)}}{\dim(\vec{x})} \quad (12)$$

$$d_{HP}(\vec{x}, \vec{y}) = \max(h'(\vec{x}, \vec{y}), h'(\vec{y}, \vec{x})) \quad (13)$$

where $dim(\vec{x})$ is the dimension/length of the compact vector $\vec{x}$. The internal distance d_e measures the difference between two values, while only distances exceeding threshold ξ (mass error tolerance) are taken into account.

4.2.4 Modifying d_{HP} by TriGen

Although d_{HP} is generally a semi-metric distance, its T-error is very low (below 0.001) but its intrinsic dimensionality is very high (above 100). Thus, we have used TriGen to improve the intrinsic dimensionality, setting the T-error tolerances to the range 0.001 – 0.2. The FP-base and 454 different RBQ-bases (different points (a, b)) were used by TriGen. Note that d_A and d_{HP} must be normalized to $\langle 0, 1 \rangle$ in order to employ the TriGen. The d_A is normalized by $\frac{\pi}{2}$, while d_{HP} by $\sqrt[n]{d_e^{max}}$, where d_e^{max} is the maximal possible value in a compact vector (i.e., the dimension of the high-dimensional representation).

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 0.2). The resulting modifiers determined by TriGen for the d_{HP} and $n = 50$ with lowest intrinsic dimensionality are shown in Tab. 1. The intrinsic dimensionality ρ using RBQ modifiers is slightly better than ρ using the FP modifier, however, testing many RBQ modifiers is time consuming.

	FP(v,w)			$RBQ_{(a,b)}$(v,w)				
T-err.tol.	ρ	T-err.	w	ρ	T-err.	a	b	w
0.001	18.6	0.001	-2.6	15.7	0.001	0.22	0.82	-3.1
0.01	7.6	0.013	-5.0	6.8	0.013	0.22	0.82	-11.3
0.02	6.0	0.023	-5.9	5.7	0.020	0.22	0.82	-20.5
0.04	4.5	0.042	-7.0	4.6	0.042	0.13	0.83	-7.6
0.06	3.8	0.062	-7.9	3.7	0.061	0.13	0.83	-10.4
0.08	3.3	0.082	-8.6	3.1	0.081	0.13	0.83	-15.3
0.1	3.0	0.102	-9.2	2.8	0.092	0.13	0.83	-20.4
0.12	2.8	0.120	-9.6	2.3	0.112	0.13	0.83	-54.9
0.14	2.6	0.138	-10.1	2.7	0.140	0.05	0.85	-6.4
0.16	2.4	0.154	-10.5	2.5	0.160	0.05	0.85	-7.1
0.18	2.3	0.173	-10.9	2.3	0.174	0.05	0.85	-7.5
0.2	2.2	0.191	-11.3	2.1	0.196	0.05	0.85	-8.4

Table 1: Empirically determined FP and RBQ modifiers and intrinsic dimensionality ρ for d_{HP}.

4.2.5 Precursor Mass Filter

As well as d_A, the d_{HP} can be extended with the precursor mass filter in order to reach much better indexability even if TriGen is not employed (14). A disadvantage is that later extension of d_{HP} for analysing of chemical modifications in the query mass spectra may be limited (Sec. 4.1).

$$d'_{HP}(\vec{x}, \vec{y}) = c_1 d_{HP}(\vec{x}, \vec{y}) + c_2 d_{mp}(m_p^x, m_p^y) \quad (14)$$

4.3 Interpretation using Similarity Search

The entire process of peptide mass spectra interpretation we propose can be summarized as follows:
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) [21]. However, even if the splitting sites are well predictable, the process is not perfect in practice and there can some missed cleavage sites 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 or Pivot tables) under d_{HP} modified by the TriGen (Sec. 4.2.4).
4) The experimental spectrum is preprocessed before 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.
5) A kNN query is processed by the MAM, while the correct peptide sequence corresponding to the spectrum could be 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 (Sec. 2).

In the experimental results we suppose 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. EXPERIMENTS

In our experiments, we used mass spectra from [10], which was obtained from 14 mass spectrometer runs on protein mixture A and 8 runs on protein mixture B. We used two query sets in our experiments. The first set Q_1 contained 119 spectra from the first run on mixture A and the second query set Q_2 contained 1941 spectra from all runs on both mixtures. The spectra split by trypsin were selected, interpreted by SEQUEST [27] and the results were manually checked by domain experts. Hence, we consider the SEQUEST-interpreted results as the confirmed ground truth.

The databases DB_1 and DB_2 were extensions of the file with correct protein sequences assigned to the mass spectra in [10]. The databases were extended with protein sequences from MSDB (release 08-31-2006) [17]. The DB_1 contained 100,000 protein sequences (5,600,747 peptide sequences) and DB_2 contained 500,000 protein sequences (29,460,880 peptide sequences).

In the experiments, the following values were measured:
a) The correctness of interpretation as a ratio of correctly assigned peptide sequences to all spectra from a query set.
b) The distance computations as the ratio of average number of distance computations per one mass spectrum interpretation to the cost of sequential scan.
c) 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 × 3GHz) with 8 GB RAM and 64-bit OS Windows Server 2008 R2. We used a C++ parallel implementation of the M-tree and the Pivot tables for indexing and querying [13]. We have also implemented parallel version of the sequential scan, as a baseline access method. Our implementation was based on Intel's Threading Building Blocks (TBB) [30]. By default, the average query time per one mass spectrum interpretation was measured on 8 processor cores.

The following settings were used unless otherwise specified – the d_{HP} was computed with $n = 50$, 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, the precursor mass error tolerance (ξ') was 2 Da, 100 peaks with the highest intensity were selected from experimental spectra. The average length of indexed vectors (the compact representation) was about 28.8 and it could be halved if just y-ions were generated instead of both y- and b-ions, however, the correctness of interpretation would be much worse.

5.1 Sequential Scan

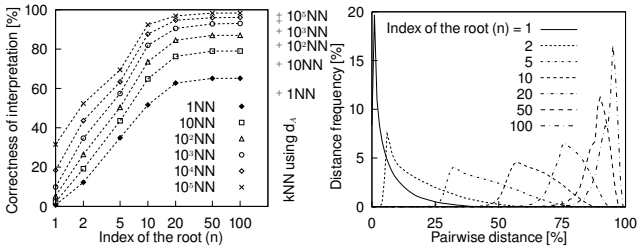

Figure 4: a) Correctness of interpretation (sequential scan), b) Distance distributions for different n.

First, the sequential scan was employed for the d_{HP} and d_A, while the correctness of interpretation and average query time were measured on DB_2 and Q_2. The correctness of interpretation is higher with increasing index of the root n (Fig. 4a). A comparison with the angle distance d_A is shown for different k in kNN queries (Fig 4a on the right). The d_{HP} has shown better correctness of interpretation than d_A. The application of the n^{th} root function in the d_{HP} has two main advantages. First, the similarity among mass spectra is modeled very well and second, the d_{HP} becomes almost a metric distance. In particular, the T-error for d_{HP} was about 0.83 for $n = 1$, but less than 0.01 for $n \geq 2$.

5.2 Improving the Indexability

A disadvantage of the n^{th} root function is that intrinsic dimensionality ρ increases with the increasing n, hence the difference between MAMs and sequential scan disappears for high n. In Fig. 4b see the distance distributions under d_{HP} (not modified by TriGen) for various n. The more the distribution is pushed to the right, the higher the intrinsic dimensionality.

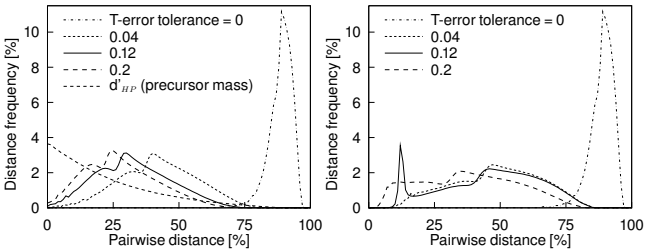

Figure 5: Distance distrib. of d_{HP} + a) FP, b) RBQ

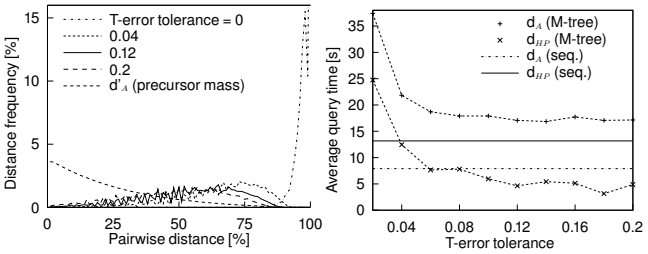

Figure 6: a) Distance distrib. of d_A + FP b) Average query time in M-tree (d_A and d_{HP}) + FP

In Fig. 5 observe the impact of T-error tolerance on the distance distributions obtained using the TriGen-modified d_{HP}, considering either FP-base or RBQ-bases. Obviously, a higher T-error tolerance leads to more convex T-modifier, and so to lower intrinsic dimensionality (distance distribution pushed to the left). In Fig. 6a, see the same for angle distance d_A, which shows its poor indexability by MAMs. In fact, about 35% of all pairwise distances are maximal $d_A = 1$ (not shown for all T-error tolerances in Fig. 6a for better readability). Note that these 35% distances cannot be fixed by the TriGen algorithm, as they are indistinguishable. In Fig. 5a and Fig. 6a, the distances d'_{HP} and d'_A (c_1=1, c_2=1) are shown for comparison (TriGen was not employed). Although theirs indexability is good, an extension of these distances for search of chemical modifications in the mass spectra may be too complicated to practical use.

The d_{HP} was compared with d_A on DB_2 indexed by M-tree, where 1000NN queries from Q_2 were used (Fig. 6b). While the d_A was 2.5× slower than sequential scan, the d_{HP} was 1.6× faster. The correctness of interpretation was 1.4× better for the d_{HP} than for d_A (compare with Fig. 7a). The average query time for d'_A and d'_{HP} was 0.4 s. The correctness of interpretation was 89.6% for d'_A and 85.7% for d'_{HP}.

5.3 Correctness of Interpretation

The following experiments were carried out on DB_2 and Q_2, while the M-tree was employed. The correctness of mass spectra interpretation is worse with increasing T-error tolerance. The kNN queries with higher k can be used to avoid this problem (Fig. 7a). The correct peptide sequences are not spread uniformly over all interval $\langle 1..k \rangle$ of a kNN query

result set but they are cumulated among a few nearest neighbors in many cases. As shown in Fig. 7b, the first nearest neighbor taken from the 100NN result was more likely to be correct than when taking the first nearest neighbor from 10NN result. The average query time of a kNN query and its distance computations ratio (wrt. sequential scan) increases with k (Fig. 8). The best results were obtained at T-error 0.06, while the correctness of interpretation was 75%, speed-up of the M-tree with respect to sequential scan was 1.7× and the distance computations ratio was 9.7%.

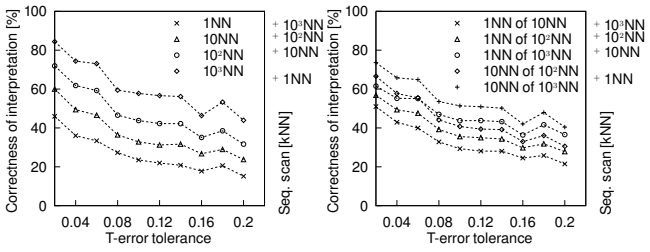

Figure 7: Correctness of interpretation (d_{HP})

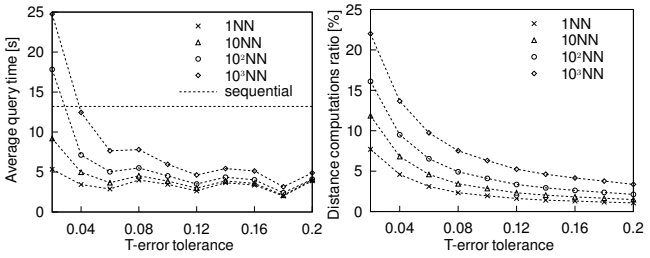

Figure 8: kNN query – a) average query time, b) distance computations ratio (wrt. sequential scan)

5.4 Comparison of M-tree and Pivot Table

We compared the efficiency of the d_{HP} (FP) indexed by M-tree and Pivot tables with the sequential scan. The experiments were made on DB_1 and Q_1, 2000NN queries were used, and 8 processor cores were employed. The Pivot table was constructed for 40 randomly selected pivots. The distance computations ratio is smaller (wrt. seq. scan) if the T-error tolerance is higher. The best results were obtained using the Pivot table for the T-error tolerance 0.04 and higher (Fig. 9b). However, the best average query time was obtained for the M-tree (Fig. 9a). Since the Pivot table was stored in the main memory, it was also fast, but the size of the Pivot table was almost 2 GB. The correctness of interpretation decreases with increasing T-error tolerance for both the M-tree and the Pivot table. (Fig. 10a).

We made comparison of the M-tree and Pivot tables on different number of processor cores (Fig. 10b) with the same settings. The M-tree on 8 cores was about 6.6× faster than M-tree on 1 core, the Pivot table was 4.9× faster and the sequential scan was 6.3× faster. If the mass spectra interpretation ran using the M-tree on 8 cores, the speed-up would be 40.1× against the sequential interpretation on 1 core.

The performance of the M-tree and Pivot tables (50 pivots) using d_{HP} was also examined on differently sized da-

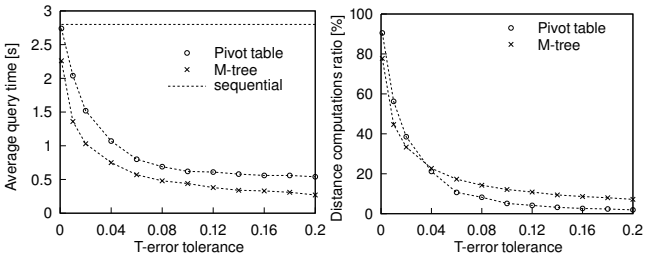

Figure 9: a) average query time, b) distance computations ratio

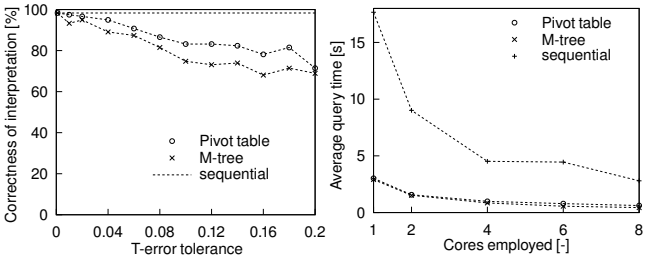

Figure 10: a) correctness of interpretation, b) number of processor cores

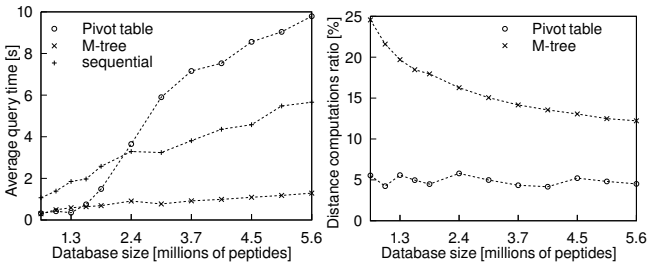

Figure 11: Database size - a) average query time, b) distance computations ratio

tabases of proteins from DB_1 (from 10 to 100 thousands of proteins; from 650 thousands to 5.6 millions of peptides or indexed vectors). The Fig. 11a shows the impact of database size on average query time, while the T-error tolerance was set to 0.1. The Pivot table is faster than M-tree as long as all its blocks are stored in main memory. If the main memory size is exceeded then Pivot table becomes inefficient. We had allocated 600 MB buffer in main memory and it was exceeded by Pivot table after 25 thousands of protein sequences (1.5 millions of peptides) were indexed. Moreover, for more than 40 thousands of protein sequences (2.4 millions of peptides) the sequential scan outperformed the Pivot tables. Fig. 11b shows that distance computations are misleading for Pivot tables when the size of main memory is exceeded.

6. 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

parametrized Hausdorff distance (d_{HP}) is a powerful combination for tandem mass spectra interpretation. The d_{HP} models the similarity among spectra very well and it can be utilized by MAMs when TriGen algorithm is employed. In general, if the T-error is higher, then indexability of the d_{HP} by MAMs is much better, the search is faster and correctness of interpretation is a little lower.

The d_{HP} models the similarity among mass spectra better than angle distance, which is commonly mentioned in mass spectrometry literature. We verified the conclusions presented in [25] that the angle distance has its limitations for utilization by MAMs in terms of mass spectra interpretation and that the angle distance in combination with the precursor mass filter is indexable very well. Moreover, we have shown that the precursor mass filter can be easily joined with d_{HP} as well as with the angle distance. A disadvantage is that combination of the angle distance or d_{HP} with the precursor mass filter might not be applicable for the query mass spectra containing chemical modifications, which is in practice a relatively frequent phenomenon. In our future work we plan to use the PM-tree for mass spectra interpretation, which could speed-up the whole task by an order of magnitude and we plan to deal with chemical modifications in the query mass spectra.

7. ACKNOWLEDGMENTS

This research has been supported in part by Czech Science Foundation project Nr. 201/09/0683 and by the grant SVV-2010-261312.

8. 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):385–387, 2004.

[2] E. Chávez and G. Navarro. A Probabilistic Spell for the Curse of Dimensionality. In *ALENEX'01, LNCS 2153*, pages 147–160. Springer, 2001.

[3] P. Ciaccia, M. Patella, and P. Zezula. M-tree: An Efficient Access Method for Similarity Search in Metric Spaces. In *Proc. of 23rd Int. Conf. on VLDB*, pages 426–435, 1997.

[4] 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*, 6(3):327–342, 1999.

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

[6] A. Frank and P. Pevzner. PepNovo: de novo peptide sequencing via probabilistic network modeling. *Analytical Chemistry*, 77(4):964–973, 2005.

[7] L. Y. Geer and et al. Open Mass Spectrometry Search Algorithm. In *Journal of Proteome Research*, volume 3, pages 958–964, 2004.

[8] D. F. Hunt and et al. Protein Sequencing by Tandem Mass Spectrometry. In *Proc. Nati. Acad. Sci. USA*, volume 83, pages 6233–6237, 1986.

[9] N. C. Jones and P. A. Pevzner. *An Introduction to Bioinformatics Algorithms*. MIT Press, Cambridge, Massachusetts, 2004.

[10] A. Keller and et al. Experimental Protein Mixture for Validating Tandem Mass Spectral Analysis. *OMICS: A Journal of Integrative Biology*, 6(2):207–212, 2002.

[11] M. Kinter and N. E. Sherman. *Protein Sequencing and Identification Using Tandem Mass Spectrometry*. John Wiley & Sons, New York, USA, 2000.

[12] J. Liu and et al. Methods for peptide identification by spectral comparison. *Proteome Science*, 5(3), 2007.

[13] J. Lokoč. Parallel dynamic batch loading in the m-tree. In *SISAP 2009, IEEE*, pages 117–123, 2009.

[14] Lutefisk. http://www.hairyfatguy.com/Lutefisk/.

[15] B. Ma and et al. PEAKS: Powerful Software for Peptide De Novo Sequencing by MS/MS. http://www.bioinformaticssolutions.com/.

[16] MASCOT. http://www.matrixscience.com/.

[17] Mass Spectrometry Protein Sequence Database. http://www.proteomics.leeds.ac.uk/bioinf/msdb.html.

[18] G. Navarro. Analyzing metric space indexes: What for? In *SISAP 2009, IEEE*, pages 3–10, 2009.

[19] J. Novák and D. Hoksza. An Application of the Metric Access Methods to the Mass Spectrometry Data. In *CIBCB 2009, IEEE*, pages 220–227, 2009.

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

[21] J. V. Olsen, S. Ong, and M. Mann. Trypsin Cleaves Exclusively C-terminal to Arginine and Lysine Residues. *Molecular and Cellular Proteomics*, 3:608–614, 2004.

[22] G. Petsko and D. Ringe. *Protein Structure and Function (Primers in Biology)*. New Science Press Ltd, London, UK, 2004.

[23] P. A. Pevzner, Z. Mulyukov, V. Dančík, 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.

[24] ProteinProspector. http://prospector.ucsf.edu/.

[25] S. R. Ramakrishnan and et al. A Fast Coarse Filtering Method for Peptide Identification by Mass Spectrometry. *Bioinformatics*, 22(12):1524–31, 2006.

[26] R. G. Sadygov and et al. Large-scale Database Searching Using Tandem Mass Spectra: Looking up the Answer in the Back of the Book. *Nature Methods*, 1(3):195–202, 2004.

[27] SEQUEST. http://fields.scripps.edu/sequest/.

[28] T. Skopal. Unified Framework for Fast Exact and Approximate Search in Dissimilarity Spaces. *ACM Transactions on Database Systems*, 32(4):29, 2007.

[29] D. L. Tabb and et al. Similarity among Tandem Mass Spectra from Proteomic Experiments: Detection, Significance and Utility. *Anal. Chem.*, 75(10), 2003.

[30] Threading Building Blocks (TBB). http://www.threadingbuildingblocks.org/.

[31] UNIMOD. http://www.unimod.org/.

[32] P. Zezula, G. Amato, V. Dohnal, and M. Batko. *Similarity Search: The Metric Space Approach (Advances in Database Systems)*. Springer, USA, 2006.