Parametrised Hausdorff Distance as a Non-Metric Similarity Model for Tandem Mass Spectrometry*

Jiří Novák and David Hoksza

Department of Software Engineering, Faculty of Mathematics and Physics,
Charles University in Prague,
Malostranské nám. 25, 118 00, Prague 1, Czech Republic
{novak, hoksza}@ksi.mff.cuni.cz

Abstract. Tandem mass spectrometry is a widely used method for protein and peptide sequences identification. Since the mass spectra contain up to 80% of noise and many other inaccuracies, there still exists a need for more accurate algorithms for mass spectra interpretation.
The sizes of protein databases grow rapidly and the methods for indexing these databases in order to interpret mass spectra become very popular. The parametrised Hausdorff distance, suitable for non-metric search, is presented in this paper. It models the similarity among tandem mass spectra very well and it is able to match the spectrum to correct peptide sequence in many cases without any post-processing scoring system.

Keywords: tandem mass spectrometry, metric access methods, peptide identification, bioinformatics

1 Introduction

Tandem mass spectrometry [8] is a fast and popular method for determining protein sequences from an experimentally prepared protein sample. Protein sequences identified by mass spectrometry are used in many fields of biological research especially in methods for protein structure and function prediction [18].

Definition 1. *Protein sequence is a linear sequence (of amino acids) over alphabet α of 20 letters, where α contains all letters from English alphabet except* $\{B, J, O, U, X, Z\}$[1].

Mass spectrometry does not determine sequences directly but the collection of data to be interpreted is obtained from tandem mass spectrometer. Each protein molecule in the sample is digested into peptides (short pieces of proteins)

* This research has been supported in part by Czech Science Foundation (GAČR) project Nr. 201/09/0683 and by institutional research plan number MSM0021620838.

[1] The omitted letters may sometimes represent more than one amino acid if there is no chance to differentiate them.

J. Pokorný, V. Snášel, K. Richta (Eds.): Dateso 2010, pp. 1–12, ISBN 978-80-7378-116-3.

by an enzyme before mass analysis. The most common and cheap enzyme is trypsin and it digests protein after each[2] amino acid K (lysine) and R (arginine) if they are not followed by P (proline) [17].

Each peptide gets a charge z in a mass spectrometer and it becomes peptide ion[3]. Peptide ions are separated by their ratio mass m (also called precursor mass) and charge z, and then they are splitted to many peptide fragment ions. The dataset obtained from the tandem mass spectrometer is a list of mass spectra (one spectrum for each detected peptide ion). Precursor mass and charge can be provided as an additional information for each spectrum corresponding to a peptide ion. The process of assigning a corresponding peptide sequence to an experimental spectrum is denoted as mass spectrum interpretation.

Definition 2. *Mass spectrum is represented by a list of peaks. Each peak corresponds to a peptide fragment ion and it is a pair of numbers m/z and intensity of occurrence, where m denotes mass in Daltons*[4] *and z charge.*

An experimentally obtained mass spectrum (acquired by division of peptide ions in mass spectrometer) usually contains many noise peaks (up to 80%), which correspond to ions with very complicated and upredictable chemical structure. The intensity may help to differentiate between more and less significant peaks in such spectrum. Spectra generated from database of protein sequences (see section 2) can be denoted as a hypothetical or theoretical. The intensity cannot be determined from sequences for peaks in a hypothetical spectrum. The hypothetical spectra do not contain intensity which usually does not cause a problem, since the m/z ratio provides the main information for mass spectra interpretation.

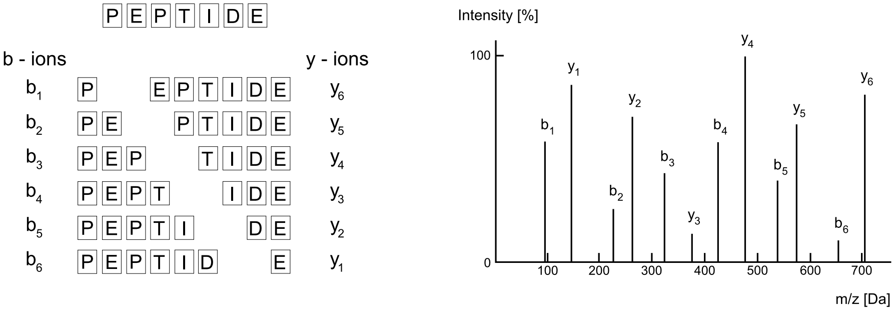

Fig. 1. Mass spectrum of sequence *PEPTIDE.*

[2] The digestion process is not perfect in practice so there can be some missed cleavage sites.

[3] Neutral molecules are not captured by mass spectrometer.

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

There are several types of fragment ions in a mass spectrum, which are fundamental for correct peptide sequence identification. The most frequently occurring are y-ions and b-ions[5] (Fig. 1). A ion serie is created by each type of ions. The completeness of y-ions or b-ions series determines the quality of the interpretation because the difference between two neighboring peaks in one serie corresponds to the mass of an amino acid. For example, missing of y_3 and b_4 in Fig. 1 causes loosing the information on the order of the letters T and I. The letters can be determined from the difference of m/z between y_2 and y_4 (or b_3 and b_5), but more candidate pairs of amino acids having similar aggregate m/z value can be selected from 20^2 possible amino acids pairs.

Modifications of amino acids are also a common problem when mass spectra are interpreted because other chemical groups can be attached to the amino acids in proteins. This usually happens during sample preparation for the mass analysis or in the mass spectrometer. The most common modifications are e.g. carbamidation of cysteine C (+57.01 Da) or oxidation of methionine M (+16 Da)[6]. The database UNIMOD [26] gathers discovered protein modifications for mass spectrometry. At the time of writing this paper, there were more than 620 known modifications.

2 State of the Art

Tandem mass spectra interpretation employs two basic approaches. *Ab initio*, the first approach, is based on direct mass spectra interpretation using graph algorithms and it is usually called as *De Novo* peptide sequencing [4]. This approach is highly influenced by occurrence of complete ion series because missed y-ions or b-ions can cause that there are many paths in graph and it is difficult to assign correct peptide sequence to the spectrum. The quality of identification using this approach is about 30% [9].

The other approach is based on search in the database [21] of already known or predicted[7] protein sequences. The hypothetical spectra of peptides are generated from database of protein sequences and compared with an experimental spectrum. A combined approach, Sequence Tag, was presented in [14]. First, a short amino acid sequence (tag) is determined by hand or by graph algorithm and then a database is searched. The most common tools for peptide identification based on searching in databases are SEQUEST [23], MASCOT [12], ProteinProspector [19], OMSSA [6], etc.

The number of data in protein databases grows exponentially every year [7] and sequential scan of the whole database becomes too slow. Modeling an index is not a trivial problem due to the noise, modifications and inaccuracies in mass spectra.

[5] Ion types are defined by the positions where splitting occurs.

[6] Special types of modifications are posttranslation modifications (PTM), which arise additionally after translation of DNA to proteins.

[7] It is possible to use raw translation of DNA sequences to protein sequences, so unknown protein sequences can be determined.

The naive method is based on indexing and querying mass spectra by their precursor mass using B-tree [2]. But there can be complications if the experimental spectrum contains modifications because m/z values of peaks and also precursor mass are shifted. The lengths of peptide sequences are usually about a few tens of amino acids. Looking for peptides with modifications can cause selection of many peptides from the database because a wide interval for precursor mass tolerance must be set up.

Several more sophisticated approaches were presented. One of them uses a suffix tree [25] for preprocessing the protein sequence database and a graph algorithm is used to preprocess tandem mass spectrum [11]. Then the suffix tree is searched against spectrum graph for candidate peptides. The correct peptide sequence is determined by a scoring function (such as HMM [27], dynamic programming [9], SEQUEST-like scoring [23], etc.).

Another method is based on using a self-organizing map (SOM) which is a type of neural network [15]. The hypothetical spectra are converted to high-dimensional vectors (Ex. 1) and then SOM is trained. The experimental spectrum is then used for a range query on SOM and the peptide candidate set is obtained and a scoring function is applied.

Example 1. Let the range of m/z values in the mass spectrum be 0-2,000 Da and let it be divided in subintervals of 0.1 Da. Each mass spectrum is then represented by a 20,000 dimensional boolean feature vector having ones at places corresponding to intervals for which m/z value in the spectrum exists.

There are also database approaches based on the properties of metric space [28]. One of them uses locality sensitive hashing in Euclidean space to preprocess peptides in the database followed by range query [5]. Another method is based on using cosine similarity and MVP-tree [20]. Using variants of cosine similarity (1) and representation of mass spectra as a high-dimensional vector (Ex. 1) is common idea in mass spectrometry literature [1].

Cosine of an angle is not a metric (see section 4) but it can be turned into metric by using arccos function. The approach based on MVP-tree [20] uses two alternatives of cosine distance. The first is called fuzzy cosine distance and it is generalization of (1). The other is called tandem cosine distance and it is combination of the fuzzy cosine distance and the precursor mass filter. Comparison of our method with this approach is presented in section 5.3.

$$\cos(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x}\boldsymbol{y}}{\|\boldsymbol{x}\| \, \|\boldsymbol{y}\|} \tag{1}$$

3 Original Idea and Improvements

3.1 Original Idea

The Hausdorff distance d_H (2) and logarithmic distance d_L (3) were proposed in [16] for tandem mass spectra interpretation. These distances describe the similarity among tandem mass spectra better than e.g. Euclidean d_E or maximum distance.

The advantage of using Hausdorff distance is that components on different positions in two vectors can be compared. The main idea of using logarithmic distance is that two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ are closer considering peptide identification 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 (Ex. 2).

$$d_H(\boldsymbol{x},\boldsymbol{y}) = \max(h(\boldsymbol{x},\boldsymbol{y}), h(\boldsymbol{y},\boldsymbol{x})), \quad h(\boldsymbol{x},\boldsymbol{y}) = \max_{x_i \in \boldsymbol{x}} \left\{ \min_{y_j \in \boldsymbol{y}} \{d_E(x_i, y_j)\} \right\} \tag{2}$$

$$d_L(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^{k} \begin{cases} \log|x_i - y_i|, & |x_i - y_i| > 1 \\ 0, & otherwise \end{cases} \tag{3}$$

Example 2. Lets assume vectors of m/z values $\boldsymbol{x} = \{148, 263, 376, 477, 574, 703\}$, $\boldsymbol{y_1} = \{148, 263, 476, 477, 574, 703\}$ and $\boldsymbol{y_2} = \{140, 270, 370, 477, 570, 710\}$. The Euclidean distance between vectors $\boldsymbol{x}$ and $\boldsymbol{y_1}$ is $d_E(\boldsymbol{x},\boldsymbol{y_1}) = 100$ and the distance between vectors $\boldsymbol{x}$ and $\boldsymbol{y_2}$ is $d_E(\boldsymbol{x},\boldsymbol{y_2}) \doteq 14.6$ but the vectors $\boldsymbol{x}$ and $\boldsymbol{y_1}$ are closer considering peptide identification. The missing number 376 in $\boldsymbol{y_1}$ means that corresponding peak in the mass spectrum is missing. On the other hand the superfluous number 476 in $\boldsymbol{y_1}$ refers to the occurrence of similar 477. The replacement of values 376 and 476 can be observed as a consequence of these inaccuracies.

The vectors of m/z values were splitted by a sliding window to many shorter vectors of constant size in order to increase quality of identification. For example a sorted vector of 12 m/z values can be generated for sequence *PEPTIDE*, these numbers correspond to y and b-ions (Fig. 1). The $(l-1) * 2 - dim + 1 = 10$ vectors must be indexed for one peptide sequence of length $l = 7$ and for vectors of dimension $dim = 3$. The short vectors were indexed by M-tree. The number of correctly assigned peptide sequences to the mass spectra was about 50-60% by using Hausdorff or logarithmic distance.

3.2 Improvements

Functions such as n^{th} root or logarithm [16] are suitable for the purpose of modeling similarity between mass spectra because these can significantly decrease an error caused by outliers.

The proposed parametrised Hausdorff distance d_{HP} (5) combine the characteristics of the n^{th} root function and Hausdorff distance, $\boldsymbol{x}$ and $\boldsymbol{y}$ are vectors of m/z values, d_E is Euclidean distance, n is index of the root and m is power modifier. The Hausdorff distance allows comparison of vectors with different sizes, which is valuable for peptide sequence identification because the mass spectra (hypothetical or experimentally obtained) have different number of peaks.

$$h(\boldsymbol{x},\boldsymbol{y}) = \frac{\sum_{x_i \in \boldsymbol{x}} \sqrt[n]{\left(\min_{y_j \in \boldsymbol{y}} \{d_E(x_i, y_j)\}\right)}}{|\boldsymbol{x}|} \tag{4}$$

$$d_{HP}(\boldsymbol{x},\boldsymbol{y}) = (\max(h(\boldsymbol{x},\boldsymbol{y}), h(\boldsymbol{y},\boldsymbol{x})))^m \tag{5}$$

Using d_{HP} noticeably increases accuracy even if no pre-processing nor post-processing algorithms are employed. Typical pre-processing algorithm is a heuristic which selects the most suitable peaks for peptide sequence identification from an experimental spectrum. The post-processing algorithm is usually represented by a scoring function which selects the best peptide sequence corresponding to an experimental spectrum from the peptide sequence candidate set obtained by an index structure.

Another improvement is significant reduction of the number of vectors that are generated from protein sequences. Only one vector of m/z values is necessary for peptide sequence representation which makes this method more usable. This was not possible in the previous version of the algorithm, since d_H and d_L required splitting in order to achieve sufficient quality of identification.

The time complexity for d_{HP} computation is $O(n^2)$ but since the lists of peaks are implicitly sorted by m/z ratio so an improvement is used and complexity $O(n)$ is achieved. The asymmetric part (see Alg. 1) of the Hausdorff distance can be computed using two nested loops. The inner loop can be broken if the minimum difference between components in two vectors is found. The position of minimum is stored and it is used as starting value for inner cycle in the next outer cycle (Alg. 1, line 3), *errTol* is mass error tolerance, *root(x,n)* is $\sqrt[n]{x}$, *power(x,m)* is x^m and *abs* computes the difference between two m/z values (in the Euclidean distance d_E).

Alg. 1. Parametrised Hausdorff Distance

```
1   float ComputeAsymmetric(sortedVector X,sortedVector Y,float errTol,float n) {
2     float sum = 0; int mem_j = 0;
3     for(int i=0;i<X.size();i++) {
4       /* position of component with minimum difference in the inner cycle
5       is >= than position in previous inner cycle */
6       min = abs(X[i]-Y[mem_j]);
7       for(int j=mem_j+1;j<Y.size();j++) {
8         if (abs(X[i]-Y[j]) < min) {
9           min = abs(X[i]-Y[j]);
10          mem_j = j;
11        }
12        /* minimum difference is achieved and better result cannot be found */
13        else break;
14      }
15      sum += (min>errTol)?root(min-errTol,n):0;
16    }
17    return sum/X.size();
18  }
19
20  float Compute(sortedVector X,sortedVector Y,float errTol,float n,float m) {
21    float left = computeAsymmetric(X,Y,errTol,n);
22    float right = computeAsymmetric(Y,X,errTol,n);
23    if (left > right) return power(left,m);
24    return power(right,m);
25  }
```

4 Metric Access Methods (MAMs)

The metric is a function that satisfies reflexivity, symmetry, non-negativity and triangle inequality [28]. Function which partially corrupts the triangle inequality is called a semimetric and the search process is denoted as non-metric [24]. The MAMs [28] were designed for fast search in databases modeled in metric spaces. The triangle inequality is crucial for organizing objects into metric regions and for pruning those regions while searching. MAM used in our experiments is a metric tree (M-tree) [3] but it can be replaced with any other MAM. MAMs support using range and k-NN (k-nearest neighbor) queries.

5 Experiments

The dataset from *Keller et al.* [10] was used in our experiments. The spectra were obtained by mixing 18 proteins together.[8] These spectra were identified by SEQUEST [23] and the results were manually checked. The spectra with charge 1^+ and 2^+, digested by trypsin and with corresponding peptide sequences contained in attached protein sequences file were selected. This file was used as a database *Keller1* containing 103 protein sequences (7,391 peptide sequences). The database *Keller2* is an extension of *Keller1* where protein sequences from MSDB (Mass Spectrometry Protein Sequence Database) [13] were added. The *Keller2* contains 10,000 protein sequences (649,481 peptide sequences). The databases and the query set from [20] were also used for comparison with cosine similarity (see section 5.3).

Following qualities were measured. The quality of identification is a ratio of correctly assigned peptide sequences to the mass spectra to all spectra from the query set (without differentiating the position of the correct peptide sequence in the obtained set). The distance computations ratio is the average number of runs of Alg. 1 per one mass spectrum to the sequential access. Since the real time is directly proportional to the distance computations ratio, we mostly use the ratio in the following text. The triangle inequality ratio is an empirically determined number of triplets of vectors satisfying the triangle inequality. The distance distribution histogram (DDH) [24] shows distribution of distances between any two vectors in the database. The distances on the x axis are normalised in order to be able to compare histograms with different values of n or m (e.g. Fig. 2b). The normalisation is possible because maximum mass of generated peptides is limited. The distance frequency is the number of pairs of vectors in the distance $d \pm \delta$ in the database, where δ is an error caused by rounding.

All experiments were carried out on a 1.6 GHz processor AMD TURION TL52 with 2 GB RAM and OS Windows XP SP2. Following settings were used unless otherwise specified - digestion enzyme: trypsin, maximum missed cleavage sites: 1, mass error tolerance: 0.4 Da, y and b-ions were generated in hypothetical spectra, 100 peaks with highest intensity were selected from experimental spectra, mass range of generated peptides: 500-5,000 Da.

[8] The 119 spectra from the first run on mixture A were used.

5.1 Index of the Root

First experiments concerned the influence of the index of n^{th} root function ($n = \{1, 2, 5, 10, 20, 50, 100\}$) on the quality of peptide sequence identification and the suitability of the parametrised Hausdorff distance for use with MAMs. Settings: $m = 1$ (modifier is off), DDHs measured on *Keller1*, quality of identification measured on *Keller2* (sequential access was employed for *Keller2*).

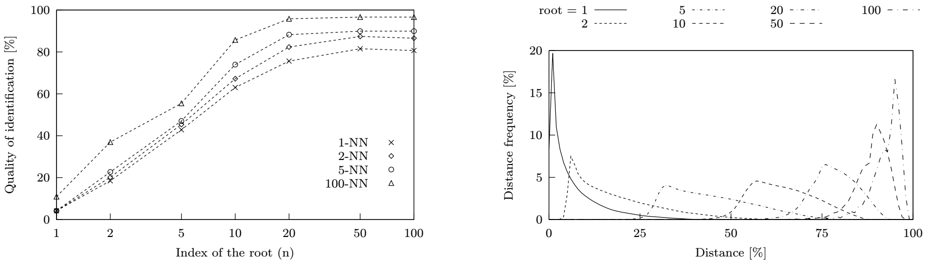

Fig. 2. Index of the root - a) quality of identification, b) DDHs.

The quality of identification increases with increasing n and the distance models the similarity among tandem mass spectra very well. The correct peptide sequences were assigned to more than 80% of experimentally obtained mass spectra as a result of 1-NN query for $n = 50$ (Fig. 2a). The number of correctly assigned sequences was about 90% for 5-NN query and more than 96% for 100-NN query. We need a 669-NN query for achieving 100% quality of identification. The selectivity is about 0.1% in such a case. The average time for the identification of one mass spectrum was about 14.4 seconds.

The triangle inequality ratio is about 17% for $n = 1$ and about 99% for $n = 2$ and higher. A disadvantage is that intrinsic dimensionality [24] gets higher with increasing n hence the distance computations ratio increases. For high n, the difference between MAMs and sequential access blends. The intrinsic dimensionality is indicated by DDH (Fig. 2b).

5.2 The Power Modifier

We tried to solve the problem of high intrinsic dimensionality by using power modifier m (5) due to poor MAMs usability. The power is monotonous function and it does not change the order of the results. The index performance (Fig. 4) was tested on M-tree with database *Keller2*.

The DDH improves with increasing modifier (Fig. 3a). Modifiers were tested for $n = 50$ (see Table 1). The DDH with $m = 1$ (modifier is off) is shown for comparison. The triangle inequality ratio gets worse with increasing power modifier (Fig. 3b). The experiments were executed for different n (see section 5.1). The quality of identification gets better with increasing triangle inequality ratio (Fig. 4a) but the distance computations ratio gets worse (Fig. 4b).

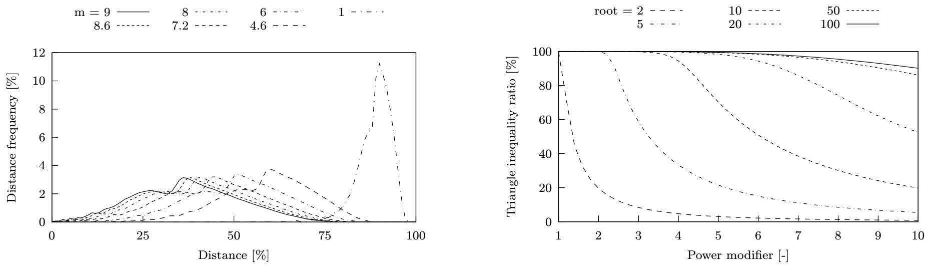

Fig. 3. The power modifier - a) corrections of DDH, b) triangle inequality ratio.

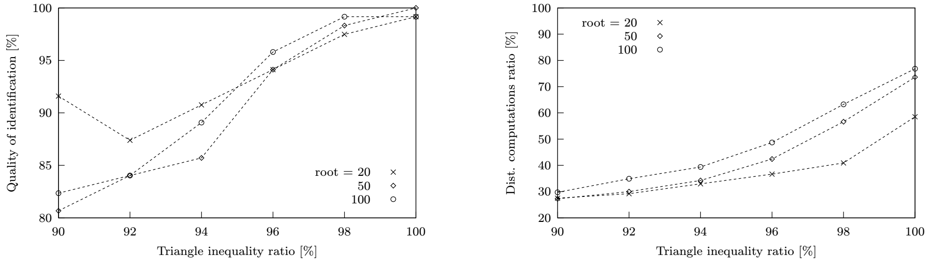

Fig. 4. The power modifier - a) quality of identification, b) dist. computations ratio.

Triangle inequality ratio [%]	90	92	94	96	98	100
$n = 20$	6.6	6.4	6	5.6	5.2	4
$n = 50$	9	8.6	8	7.2	6	4.6
$n = 100$	10	9.2	8.6	7.6	6.2	4.8

Table 1. Empirically determined modifiers m.

5.3 Comparison with the Cosine Similarity

Parametrised Hausdorff distance was compared with fuzzy cosine distance and tandem cosine distance described in [20]. Datasets described in Table 1 in the cited paper were used for the comparison. The database I contains 92,768 hypothetical spectra from the genome of Escherichia Coli K12 and 7 proteins mixture from Sashimi proteomics repository [22]. The database II has 654,276 spectra and it is an extension of database I containing hypothetical spectra from human genome. The query set contains 49 experimentally obtained spectra and comes from the 7 proteins mixture from Sashimi proteomics repository. The following settings were used: $n = 1000$, $m = 4$, 0 missed cleavage sites, error 1.0 Da, y and b ions were generated in hypothetical spectra, 100 peaks with highest intensity were selected from experimental spectra, peptide mass range 0-5,000 Da. We used 13-NN query on M-tree and the triangle inequality ratio was 99.9%.

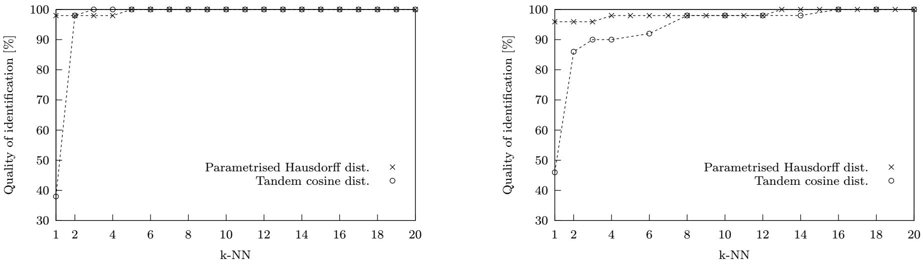

Fig. 5. Quality of identification - a) database I, b) database II.

The parametrised Hausdorff distance returns peptide sequence corresponding to the experimental spectrum as a result of 1-NN query in 98% on database I and in 95.9% on database II. The quality of identification eliminates the need of a scoring system (see section 2). But in fact the quality decreases with increasing database size and the scoring system cannot be completely removed from a real-world application.

The parametrised Hausdorff distance has better quality of identification than tandem cosine distance (Fig. 5)[9]. But in fact the tandem cosine distance has the distance computations ratio less than 0.3% for both databases and parametrised Hausdorff distance has the computations ratio 62.3% for the database I and 50.7% for the database II. Although the computations ratio of our method decreases with increasing database size, it is still slower than the tandem cosine distance. Fuzzy cosine distance has the distance computations ratio about 95%. Tandem cosine distance's computation ratio is a consequence of combination fuzzy cosine distance and precursor mass filter [20]. The precursor mass filter can be restrictive criterion if the peptide modifications are searched. Typical precursor mass tolerance is about ±2 Da. This tolerance must be extended for searching peptides with modifications. Precursor mass of modified peptides can differ by more than a few tens to hundreds Daltons.

5.4 Non-Metric Search and k-NN Queries

An interesting characteristic can be observed when non-metric search is used. We examined the performance of the M-tree (*Keller2*) using $n = 50$ and $m = 9$ which corresponds to 90% triangle inequality ratio. The k in k-NN query was increased and the quality of identification grew. The results were not distributed uniformly over the interval of k items but the correct peptide sequences were found as the top hits in many cases (Fig. 6a). This is a consequence of non-metricity and it cannot happen if the distance is metric or if the sequential access is used. The distance computations ratio and average time of identification per one spectrum grew with increasing k in k-NN query (Fig. 6b). Average time was about 15.2 seconds for sequential access.

[9] The results for tandem cosine distance were taken from the supplement of [20].

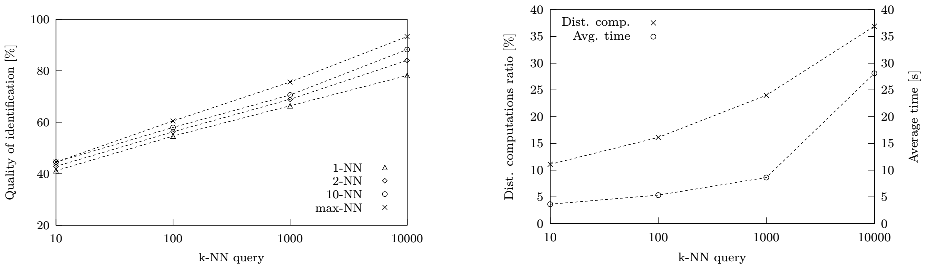

Fig. 6. Non-metric search and k-NN queries - a) quality of identification, b) distance computations ratio and average time.

6 Conclusions and Future Work

The parametrised Hausdorff distance for interpretation tandem mass spectra of peptides was proposed. It was compared with cosine distance which is widely discussed in mass spectrometry literature. The fuzzy and tandem cosine distance were used in this paper. Tandem cosine distance shows worse results in terms of quality identification than our algorithm. The fuzzy approach is moreover slower in terms of distance computations ratio. Higher speed of tandem cosine distance is a consequence of including the precursor mass filter. On the other hand, embedding of precursor mass filter can be problematic when modeling of the similarity of spectra corresponding to modified peptides is desired. Development of more precise semimetrics can also reduce the need of complicated scoring algorithms. The design of better modifier functions for parametrised Hausdorff distance opens possibilities for further research. Finally, the abilities of k-NN query for non-metric search were presented.

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*, vol. 15, issue 3, pp. 385-387. 2004.
2. R. Bayer and E.M. McCreight. Organization and Maintenance of Large Ordered Indices. *Acta Inf.*, vol. 1, pp. 173-189. 1972.
3. P. Ciaccia, M. Patella and P. Zezula. M-tree: An Efficient Access Method for Similarity Search in Metric Spaces. *Proc. of 23rd Int. Conf. on VLDB*, pp. 426-435. 1997.
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*, vol. 6, no. 3, pp. 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*, vol. 23, no. 5, pp. 612-618. 2007.

6. L.Y. Geer et al. Open Mass Spectrometry Search Algorithm. *Journal of Proteome Research*, vol. 3, pp. 958-964. 2004.
7. D. Hoksza and T. Skopal. Index-based approach to similarity search in protein and nucleotide databases. *CEUR Proc. Dateso 2007*, vol. 235, pp. 67-80. 2007.
8. D.F. Hunt et al. Protein sequencing by tandem mass spectrometry. *Proc. Nati. Acad. Sci. USA*, vol. 83, pp. 6233-6237. 1986.
9. N.C. Jones and P.A. Pevzner. *An Introduction to Bioinformatics Algorithms.* MIT Press, Cambridge, Massachusetts. 2004.
10. A. Keller et al. Experimental Protein Mixture for Validating Tandem Mass Spectral Analysis. *Journal of Integrative Biology*, vol. 6, no. 2, pp. 207-212. 2002.
11. B. Lu and T. Chen. A suffix tree approach to the interpretation of tandem mass spectra: applications to peptides of non-specific digestion and post-translational modifications. *Bioinformatics Oxford Journal*, vol. 19 (Suppl. 2), pp. 113-121. 2003.
12. MASCOT. `http://www.matrixscience.com/`.
13. Mass Spectrometry Protein Sequence Database (MSDB). `http://www.proteomics.leeds.ac.uk/bioinf/msdb.html`.
14. E. Mortz et al. Sequence tag identification of intact proteins by matching tandem mass spectral data against sequence data bases. *Proc. Natl. Acad. Sci. USA*, vol. 93, pp. 8264-8267. 1996.
15. K. Ning, H.K. Ng and H.W. Leong. PepSOM: An Algorithm for Peptide Identification by Tandem Mass Spectrometry based on SOM. *Genome Informatics*, vol. 17, pp. 194-205. 2006.
16. J. Novák and D. Hoksza. An Application of the Metric Access Methods to the Mass Spectrometry Data. *IEEE CIBCB 2009.* Nashville, TN, USA. ISBN 978-1-4244-2756-7, pp. 220-227.
17. J.V. Olsen, S. Ong and M. Mann. Trypsin Cleaves Exclusively C-terminal to Arginine and Lysine Residues. *Molecular and Cellular Proteomics*, vol. 3, pp. 608-614. 2004.
18. G.A. Petsko and D. Ringe. *Protein Structure and Function (Primers in Biology).* New Science Press Ltd, London, UK. 2004.
19. ProteinProspector. `http://prospector.ucsf.edu/`.
20. S.R. Ramakrishnan et al. A fast coarse filtering method for peptide identification by mass spectrometry. *Bioinformatics Oxford Journal*, vol. 22, no. 12, pp. 1524-1531. 2006.
21. R.G. Sadygov, D. Cociorva and J.R. Yates III. Large-scale database searching using tandem mass spectra: Looking up the answer in the back of the book. *Nature Methods*, vol. 1, no. 3, pp. 195-202. 2004.
22. Sashimi proteomics repository. `http://sashimi.sourceforge.net/repository.html`.
23. SEQUEST. `http://fields.scripps.edu/sequest/`.
24. T. Skopal. Unified Framework for Fast Exact and Approximate Search in Dissimilarity Spaces. *ACM Transactions on Database Systems (TODS)*, vol. 32, issue 4. 2007.
25. E. Ukkonen. On-line construction of suffix trees. *Algorithmica*, vol. 14, pp. 249-260. 1995.
26. UNIMOD. `http://www.unimod.org/`.
27. Y. Wan, A. Yang and T. Chen. PepHMM: A Hidden Markov Model Based Scoring Function for Mass Spectrometry Database Search. *Anal. Chem.*, vol. 78, pp. 432-437. 2006.
28. P. Zezula, G. Amato, V. Dohnal and M. Batko. *Similarity Search: The Metric Space Approach (Advances in Database Systems).* Springer, New York, USA. 2006.